Title

Examining the Effects of Linguistic Complexity on Emergent Bilinguals' Academic Content Performance

Author

Rowe, Susan E.

Publication Date

2023

Examining the Effects of Linguistic Complexity on Emergent Bilinguals' Academic Content Performance

By

SUSAN ROWE
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Education

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____
Megan E. Welsh, Chair

____
Anthony D. Albano

____
Nicole Sparapani

Committee in Charge

2023

Abstract

This dissertation explored whether unnecessary linguistic complexity (LC) in mathematics and biology assessment items changes the direction and significance of differential item functioning (DIF) between subgroups emergent bilinguals (EBs) and English proficient students (EPs). Due to inconsistencies in measuring LC in items, Study One adapted a rubric counting instances of specific grammatical features in items and introduced a method for evaluating lexical features in items. Four raters were asked to count the presence of five grammatical features in assessment items and determine whether each feature contained construct-relevant vocabulary. The items were drawn from four content assessments administered to Massachusetts high school students: two biology assessments and two mathematics assessments. These counts of grammatical and lexical features were modeled in factor analyses to evaluate the multidimensionality of LC and subsequent fit of multidimensional LC models. While there were problems with raters consistently counting construct-irrelevant grammatical features, multidimensional models of LC fit acceptably well. Factor scores obtained from the measurement models for lexical complexity, relative clauses, and complex noun phrases created in Study One were used for Study Two.

In Study Two, Rasch hierarchical generalized linear models (HGLMs) were created to evaluate DIF between different subgroups of EBs and EPs on a biology assessment and a mathematics assessment, as including LC as an item covariate may predict item responses differently by comparison group. Seven comparison groups were evaluated across two assessments (mathematics and biology): EPs versus EBs, EPs versus short-term EBs, EPs versus long-term EBs, short-term EBs versus long-term EBs, EPs versus Spanish-speaking EBs, EPs versus non-Spanish-speaking EBs, and non-Spanish-speaking EBs versus Spanish-speaking EBs

(reference group versus focal group, respectively). For each comparison group, at least five models were created: a comparison model with all participants in the comparison group with that only accounts for the main effect of focal group status, a “base model” that evaluated DIF for the comparison groups with no LC item covariates, a model including lexical complexity as an item covariate (“LEX predictor”), a model including complex noun phrases as an item covariate (“NP predictor”), and a model including relative clauses as an item covariate (“RC predictor”). If LC predictor models improved model fit, models with multiple LC predictors were created.

For the EP versus EB comparison groups on the mathematics assessment, model fit only improved with the NP predictor model, while the LEX, NP, and RC predictor models improved model fit for the EB versus EB comparison groups; a model with all LC predictors improved model fit for the EB versus EB comparison groups. For the biology assessment, the LEX, NP, and RC predictor models improved model fit for all comparison groups; a model with all LC predictors improved model fit for all comparison groups. The main effects of the item covariates (LC factor scores) and their interactions with focal group status were evaluated, as were the number of items within a comparison group that had changes in DIF significance or direction when including a LC predictor. All LC predictors had consistent main effects across comparison groups. For the mathematics assessment, items with higher complex noun phrases factor scores were consistently more difficult for all comparison groups (NP predictor model), and items with higher lexical complexity (LEX predictor model, all predictors model) or relative clauses factor scores (RC predictor model, all predictors model) were consistently more difficult for all EB versus EB comparison groups. For the biology assessment and all comparison groups, items with higher lexical complexity (LEX predictor model, all predictors model) or complex noun phrases factor scores (NP predictor model, all predictors model) were consistently more difficult, and items with lower relative clauses factor scores (RC predictor model, all predictors model) were consistently more difficult, with one exception. In the all predictors models for the EB versus EB comparison groups, only relative clauses had a significant main effect.

There were some changes in interactions with LC predictors and focal group status. For the mathematics assessment and EP versus EB comparison groups, complex noun phrases interactions favored EPs. For the mathematics assessment and EB versus EB comparison groups, generally the interactions in the single LC predictor models generally favored STEBs compared to LTEBs and non-Spanish-speaking EBs compared to Spanish-speaking EBs, but when all LC predictors were included, no interactions between LC predictor and focal group status were significant. For the biology assessment and EP versus EB comparison groups, lexical complexity and complex noun phrases factor scores interactions generally favored EPs, and relative clauses factor scores interactions favored EBs and EB subgroups. For the biology assessment and EB versus EB comparison groups, regardless of whether examining the single LC predictor or all predictors models, no interactions between focal group status and LC predictor were significant.

Changes in DIF significance and direction were compared between the base model and LC predictor models for all comparison groups. For the mathematics assessment and EP versus EB comparison groups, after conditioning on complex noun phrases, items with complex noun phrases generally exhibited significant DIF favoring EBs, regardless of whether the complex noun phrases factor scores were high (one standard deviation above the mean) or low (due to floor effects, the lowest complex noun phrases factor score). For the biology assessment, all items exhibited significant DIF favoring EBs after accounting for lexical complexity, most items exhibited non-significant DIF after accounting for complex noun phrases or relative clauses, and items were mixed between exhibiting non-significant DIF or significant DIF favoring EBs after accounting for all LC predictors. While items with high relative clauses factor scores exhibited non-significant DIF, some items with low relative clauses factor scores exhibited significant DIF favoring EPs after accounting for relative clauses. Items with two or more high factor scores exhibited non-significant DIF, but items with two or more low factor scores exhibited significant DIF favoring EBs after accounting for all LC predictors. These results were fairly consistent across different EP versus EB comparison groups, although different items were flagged for DIF in initial models not accounting for LC predictors. Items were less difficult for EBs than EPs after accounting for LC features, which suggests the abilities of EBs are underestimated due to LC in items, even if the items have low LC. Considering subgroup differences in these EIRMs, the key takeaway is that while different items are flagged as exhibiting significant DIF for different EP versus EB comparison groups when examining DIF with no LC predictors, there are few subgroup differences in items changing DIF significance or direction after accounting for LC predictors.

Acknowledgements

I would like to express my deepest gratitude to Megan Welsh for supporting me and offering a space in your lab for me. You have been a constant source of support, I appreciate our GAANN group. Thank you for being my chair when I needed it. Jamal Abedi, thank you for your guidance and our conversations on validly assessing EBs, you are missed. I am also thankful for Tony Albano and Nicole Sparapani; thank you for serving on my committee. Your feedback was helpful. This endeavor would not have been possible without the support from the School of Education, who financed my research.

Lastly, I'd like to thank my family. Sean, thank you for being my anchor while I finished this journey, I'm so excited for this next chapter in our life together. Thank you to my mom, sister, and cousins; you've always believed I could do it!

Dedication

This dissertation is dedicated to my husband, who has been a source of support and comfort throughout this project.

Table of Contents

CHAPTER ONE

Introduction

When testing students to determine their mastery of content, assessment developers aim to produce scores free from test bias so valid interpretations about test-takers' content mastery can be made. Test bias refers to systematic errors in the calculation of test scores by group, leading to an unfair assessment (Zieky, 2015). Assessment developers can evaluate the presence of possible test bias in the development of assessments in the item writing process by using methods including item screening by content experts or conducting think-a-loud protocols by asking test-takers to describe their thinking as they are presented with the item. While these techniques are useful in developing assessments with reduced test bias, the use of psychometrics is used to identify potential sources of bias after assessments have been taken by test-takers, most commonly through the use of differential item functioning or DIF, derived from item response theory (IRT), a latent trait score theory (Bandalos, 2019).

Unnecessary linguistic complexity (LC) in assessment items has been identified as a potential source of error in the assessment of emergent bilinguals (EBs) (Abedi et al., 1997). LC may influence DIF because EBs have more difficulties with reading comprehension in English, particularly with the linguistically complex language present on large-scale assessments. Much of the research looking at the effect of LC on DIF between EBs and non-EBs has focused on individual linguistic features such as passive voice, complex verbs, subordinate clauses, relative clauses, and noun phrases (Banks et al., 2016; Haag et al., 2013; Heppt, et al., 2015; Kachchaf et al., 2016; Shaftel et al., 2006; Turkan & Liu, 2012). However, Martiniello (2009) synthesized many of these studies and concluded a more holistic approach to measuring LC is needed due to inconsistencies in prior research on the effect of individual LC features on item responses, although others have partitioned LC into lexical and grammatical complexity (Avenia-Tapper & Llosa, 2015; Lee & Randall, 2011; Wolf & Leon, 2009). Lexical and grammatical features are distinct from each other; lexical feature operate at the word-level and grammatical features operate at the sentence-level. The multidimensionality of LC is examined in this dissertation by creating a multidimensional model of LC partitioned into factors for lexical complexity and grammatical features. Factor scores from models for lexical complexity and specific grammatical features were inserted into Rasch hierarchical generalized linear models for two assessments to evaluate the effect of LC on DIF between EBs and non-EBs, non-EBs and subgroups of EBs, and subgroups of EBs.

Emergent Bilinguals: Moving Away from a Deficit Label

This dissertation uses the term “emergent bilinguals” to refer to students formally classified as not having “sufficient” understanding of English to learn in mainstream classrooms without language support; examples of these students include recent immigrants to the United States and students born in the United States to families speaking a language other than English at home. “Emergent bilinguals” is a relatively new term for EBs introduced by Garcia et al. in 2008. Historically, these students have also been known as “English language learners,” “English as a second language,” or “limited English proficient;” EBs are labeled as such because they are deemed as having “insufficient” proficiency in English to receive instruction in a mainstream English-speaking classroom without language support although these students are learning and gaining proficiency in multiple languages (García et al., 2008). These past labels centered on a student’s proficiency in English to identify them rather than focus on their bilingualism. There are many benefits to bilingualism and it is unnecessary to use a deficit label to refer to these capable and competent learners even if they have not fully mastered the dominant language.

EB status is tracked in order to identify EBs when they enter elementary or secondary education to when they are "reclassified" as English proficient, or meeting state requirements for being considered fluent enough in English to meet education standards in English. When an EB has met the criteria of their state to be reclassified as "English proficient," they are removed from language support programs and placed into classrooms. It is up to individual states in the United States to determine the criteria for reclassification, but many follow similar criteria as suggested by García et al. (2008). To attain reclassification, EBs in Massachusetts, a state with an average population of EBs compared to the United States national average, the population sampled in the present dissertation, must meet certain thresholds for overall score and literacy composite score on ACCESS for ELLs, an English language proficiency assessment, and receive their teachers' recommendation to be reclassified. Teachers use school grades, teacher observations, and MCAS results (the state's standards-based achievement tests) to determine their reclassification recommendations (DESE, 2022). While EB is viewed as a binary label, the acquisition of language is on a continuum; there are many factors that contribute to the varying English proficiency of EBs (Solano-Flores, 2014).

EBs are a heterogenous population with varying characteristics that contribute to their English proficiency. Length of time as a EB matters, particularly in later grades, where students who have been EBs for six or more years may need different instructional approaches or language support compared to recent immigrants. Recent immigrants have varied educational backgrounds; some recent immigrants may have had formal educational experiences in their country of origin while other students may have had limited or interrupted formal education. The languages and dialects of EBs also matters. The majority of EBs are Spanish-speakers and this may overshadow EBs speaking other languages. Studies by Solano-Flores and Li have found

Spanish-speaking, Haitian-Creole, and Chinese language speaking EBs' assessment performance appears to depend on their strengths and weaknesses in their native language and English and the linguistic challenges of items given in their native language and English (Solano-Flores, 2014; Solano-Flores & Li, 2009; Solano-Flores & Li, 2006). All these varying factors contribute to variations in English proficiency, which may influence the item responses of EBs and their subsequent performance.

Linguistic Complexity and Test Bias

Many researchers have examined whether reducing unnecessary LC on assessment items by modifying assessment items can improve the accuracy of EBs' scores, reducing the construct-irrelevant variance associated with English proficiency, without unduly influencing the scores of non-EBs. Some studies have found EBs have scored higher on linguistically modified assessments than on unmodified assessments, with no differences in performance on the two types of assessments for non-EBs, which suggests LC may be unfairly influencing assessment performance for EBs (Abedi & Lord, 2001; Haag et al., 2015; Sato et al., 2010).

When an item is measuring a construct instead of, or in addition to, the construct intended to be measured, the item is said to have construct-irrelevant variance, or CIV (Abedi, 2015; Haladyna & Downing, 2004; Messick, 1989). CIV is anything influencing test-takers' scores unrelated to the measured construct (Abedi, 2002; Haladyna & Downing, 2004; Messick, 1989; Young, 2008). CIV is particularly problematic when it influences one group of test-takers over another, resulting in inaccurate assessment performance comparisons between the two groups. Unnecessary LC in items has been identified as one potential factor contributing to CIV in items. For example, if unnecessary LC in an item measuring a targeted construct affects EBs but not non-EBs, then the item may be measuring only the targeted construct for non-EBs, but the targeted construct and English proficiency for EBs if unnecessary LC is introduced into the item. However, we can determine what items may unfairly favor one group over another through DIF. As discussed previously, DIF analyses evaluate whether test-takers from different groups with the same ability have different probabilities of responding correctly to an item (Bandalos, 2019).

Many studies have looked at whether LC is predictive of DIF between EBs and non-EBs. By correlating DIF with specific linguistic features, Heppt et al. (2015) found significant correlations between linguistic features and DIF against EBs (favoring non-EBs), suggesting there is an influence on increased LC on the effect of DIF. However, Lee and Randall (2011) found math assessment items with higher LC (up to a particular point) were better indicators of math ability for non-EBs than for EBs, with many items identified as having DIF favoring the responses of non-EBs compared to EBs, although LC did not predict the effect size of DIF. Turkan and Liu found mixed results in their DIF analysis of a science assessment - three items favoring the responses of non-EBs compared to EBs, but one item favored EBs (2012). A more in-depth review of the relationship between LC and DIF can be found in Chapter Two.

Due to differences in how LC is defined by each study, it remains unclear what unnecessarily complex linguistic features may inadvertently influence differences in item responses between EBs and non-EBs, let alone how these affect subgroups of EBs, which are not evaluated in DIF research, although Lane and Leventhal (2014) advocate for routine evaluations of DIF in subgroups of EBs. There is a need for a more systematic method of identifying LC; as LC is often evaluated by human raters, standardizing the way LC is operationalized and evaluated can lead to more consistent findings about the effects of LC on DIF and help reduce measurement error, leading to more substantiated conclusions about the effects of LC on EBs' item-level responses. LC must be clearly operationalized before examining the relationship between LC and EBs' item-level responses. Past studies on the effects of LC on EBs' assessment performance vary widely with how LC was defined. Some researchers examined LC through specific linguistic features (Banks et al., 2016; Haag et al., 2013; Heppt, et al., 2015; Kachchaf et al., 2016; Shaftel et al., 2006; Turkan & Liu, 2012), a composite of LC (Mahoney, 2008; Martiniello, 2009), or subcategories of LC comprised of similar linguistic features (Avenia-Tapper & Llosa, 2015; Lee & Randall, 2011; Wolf & Leon, 2009). Studies also use different DIF analytical techniques. Most studies correlate DIF with LC (Kachchaf et al., 2016; Heppt et al., 2015; Haag et al., 2013), others have examined differential bundle functioning in items bundled by degree of LC (Banks et al., 2016; Wolf & Leon, 2009), and others have screened LC features in items exhibiting DIF (Martiniello, 2008).

Scope of Dissertation

There appear to be mixed results on what specific linguistic features (or combination of features) contribute to unnecessary LC in assessment items that unfairly affect EBs over non-EBs, as well as what EB subgroup characteristics may influence the effect of unnecessary LC on EBs. Using a method of measuring LC split into lexical and grammatical complexity may better illustrate the cumulative effects of LC on EB assessment performance. However, a comprehensive instrument measuring lexical and grammatical complexity in assessment items has yet to be developed for use in measuring construct-irrelevant LC that may influence the item responses of EBs. Such an instrument would need to evaluate whether the lexical and grammatical features identified are construct-relevant, as the presence of more complex lexical and grammatical features is not necessarily irrelevant to the measured construct (Avenia-Tapper & Llosa, 2015). The language that is assessed and is expected to be understood to demonstrate proficiency in the assessed construct is construct-relevant language (Olivieri, 2019). For example, the impact of scientific vocabulary on a biology test may be greater for EBs than for their English-proficient peers, but just because this impact exists does not mean it is a source of item bias or construct-irrelevant variance that is unfairly introduced to the construct. Construct-irrelevant language is language that is not needed to be understood to answer the item correctly, and by evaluating the linguistic complexity of items, construct-relevant and construct-irrelevant language can be disentangled (Abedi, 2015).

Study One

When it comes to designing or modifying an instrument, the reliability and validity of the measure should be established before use, including the consistency of ratings provided by trained raters. Therefore, my first study (Chapter Two) utilized a generalizability theory decision study, based on classical test theory, to determine how many raters are needed to have a reliable and valid counts of five grammatical features (passive voice, complex verbs, subordinate clauses, relative clauses, and complex noun phrases) for high school biology and mathematics assessments in one state. Raters were also asked to evaluate whether each grammatical feature identified contained construct-relevant vocabulary. While items found to exhibit DIF are presumed to have CIV, when evaluating the cause of DIF, researchers need to consider whether the differences between groups are actually caused by CIV. If there are group differences between EBs and non-EBs because of item length, general academic vocabulary used, or lengthy words, the presence of DIF may indicate bias introduced by CIV, as these are factors unrelated to the construct measured on an assessment. However, if there are group differences between EBs and non-EBs because of technical vocabulary used or other constructs intended to be measured by the assessment, this may not be an issue of test bias, but an issue of access to taught content. Therefore, the construct relevance of the language used in assessments must be considered when determining a source of DIF in items (Avenia-Tapper & Llosa, 2015). Therefore, grammatical features with construct-relevant vocabulary (biology vocabulary on the biology assessments and mathematics vocabulary on the mathematics assessments), were not included when creating the models for grammatical features associated with LC.

The measurement of LC is made up of many factors, and partitioning the measurement error with generalizability theory is a promising way to untangle this source of error (Solano-Flores, 2014; Solano-Flores et al., 2014). After identifying the sources in error in counting construct-irrelevant counts of features, I examined the multidimensionality of LC by conducting confirmatory factor analyses combining the grammatical features counted by raters with an instrument for counting lexical features (total words, general academic vocabulary, technical vocabulary, and long words). These counts of grammatical and lexical features were used to examine whether multidimensional models of LC fit the data better than unidimensional model of LC. After selecting two well-fitting factor analysis models of LC (one for the biology assessments and one for the mathematics assessments), factor scores for lexical complexity and certain grammatical feature counts (relative clauses and complex noun phrases) were obtained. These factor scores were used to evaluate the effect of different aspects of LC on item responses for Study Two.

The key research questions for Study One are as follows:

1. How many raters are needed to reliably estimate the presence of five grammatical features in assessment items?
2. What contributions do lexical features make to a lexical complexity factor score? What contributions do grammatical features make to a grammatical complexity factor score? What

contributions do lexical complexity and grammatical complexity factors make to a LC factor score? Is LC measured this way multidimensional?

Study Two

Study Two (Chapter Three) evaluated how LC factor scores influenced what items are flagged for DIF between different subgroup comparisons of EBs and non-EBs. This was accomplished by using explanatory item response theory (EIRM), an extension of item response theory (IRT) and applying it to a Rasch hierarchical generalized linear model (HGLM). EIRM uses nonlinear mixed models (such as Rasch HGLMs) to model item responses within persons. Lexical complexity, complex noun phrases, and relative clauses factor scores obtained from Study One were included as item-level covariates. The effect of LC features and their interactions with EB status were analyzed, along with which items changed DIF significance and direction when LC features were accounted for in EIRMs.

Solano-Flores (2014) discusses past research utilizing generalizability (G) theory to partition variability in item performance across students nested within language (English as a first or second language) and how the interaction of students, items, and language was found to be the largest source of error, suggesting the characteristics of students influence their assessment performance based on item-level and test-level contexts. Student responses to items are shaped by the languages and dialects they speak (including English as a first or second language) and how well they understand these languages and dialects represented in the items. Yet, specific characteristics of EBs are not taken into account when evaluating the item responses of EBs compared to non-EBs. Few studies have reported the specific characteristics of the EBs in their sample or explored the differences between EB subgroups' item responses, calling into question which EBs may be more heavily influenced by the bias introduced by LC on content assessments. EBs are a heterogenous population with a variety of characteristics that influence English proficiency and access to taught content, such as length of time as an EB or native language spoken. With a greater understanding of which EB characteristics interact with item-level linguistic features, we can better design not only assessment items but more differentiated instruction for EBs. In Study Two, DIF analyses were conducted for different comparison groups of EBs (based on length of time as an EB or native language spoken) versus non-EBs. DIF analyses were also conducted between EBs based on their length of time as an EB or native language spoken.

The key research questions for Study Two are as follows:

3. How does linguistic complexity of the test item affect item difficulty for EBs compared to non-EBs on content assessments?
4. Does accounting for linguistic complexity lead to differences in uniform DIF significance or direction when evaluating DIF between EBs and non-EBs?
5. Which EB subgroups exhibit differential functioning? Are there differences by subgroups of EBs in how accounting for linguistic complexity affects uniform DIF significance or direction?

Study Conclusions and Dissertation Format

The two studies described above work together to examine the effects of construct-irrelevant LC on EBs. Study One provided evidence as to how consistently grammatical features can be counted by trained raters, along with clear definitions for how to count specific grammatical and lexical features predicted to introduce LC in assessment items that may affect EB assessment performance. This study also included a factor analysis to determine whether LC was multidimensional, whether grammatical features contributed to a multidimensional model of grammatical complexity, and how lexical features contribute to lexical complexity. Study Two is a rare study of exploring a potential source of DIF between EBs and non-EBs by including LC as an item covariate into an IRT model and identifying how accounting for LC changes DIF significance and direction. This study also examined the differences in how accounting for LC changes DIF significance and direction between EB subpopulations. Some subpopulations of EBs may have their assessment performance affected differently than other EBs. Study Two is a novel application of differential functioning analyses in general because the heterogeneity of EBs is simply not accounted for in performance differences in different types of EBs, let alone differential functioning. Different groups of EBs have different needs and if differences in their assessment performances are identified, it may serve as evidence for instructional change for these learners or different considerations need to be made when assessing subpopulations of EBs (Lane & Leventhal, 2015).

In the present chapter (Chapter One), I discussed the context for the studies I conducted along with my research questions. Following this introductory chapter, each study will be presented in its own individual chapter. Chapter Two will present Study One ("Measuring Linguistic Complexity in Assessment Items for Emergent Bilinguals Using Generalizability Theory") and Chapter Three will present Study Two ("The Effects of Linguistic Complexity on Item Bias Against Emergent Bilinguals: An Explanatory IRT Approach"). While the studies were designed together, each study answers different research questions and therefore each chapter has its own literature review, methodology, results, and discussion sections, unlike a traditional dissertation format. The last chapter of this dissertation (Chapter Four) synthesizes the results of both studies.

CHAPTER TWO

Measuring Linguistic Complexity in Assessment Items for Emergent Bilinguals Using Generalizability Theory

Measuring Linguistic Complexity

In this section, I will discuss past efforts in measuring LC thought to influence EB item responses. First, I will discuss general approaches to measuring LC and which of these approaches may best capture the effect of LC on assessment performance. Then I will discuss the research conducted around individual linguistic features and their noted effects on DIF between EBs and non-EBs. Research thus far indicates inconsistencies in how LC is operationalized across studies and evaluating the effects of LC on assessment performance by individual linguistic features may lead to these inconsistent results.

Researchers have found limited effects for targeting specific linguistic features to reduce potential bias in items, but few have examined whether evaluating LC holistically or aggregating by type of linguistic feature predicts DIF or item difficulty. Martiniello (2009) noted inconsistencies in which linguistic features predicted differences in item difficulty between EBs and non-EBs and concluded LC needs to be scored as a composite of overall LC, as previous studies looking at individual linguistic features and aggregating by category did not have consistent findings on how LC affects item difficulty. Other researchers examined the effects of a lexical complexity composite and a grammatical complexity composite (Avenia-Tapper & Llosa, 2015; Lee & Randall, 2011). Avenia-Tapper and Llosa (2015) note while lexical complexity (word-based linguistic features such as uncommon vocabulary) is well-studied in EB assessment research, grammatical complexity (sentence-level linguistic features such as subordinate clauses) is less systematically studied. Lee and Randall (2011) did rate the lexical and grammatical complexity of items, but found limited effects of LC on EB and non-EB item responses, although the authors noted the math assessment used in their study had low levels of lexical and grammatical complexity and more ratings from linguistic experts were needed to improve reliability.

Despite the amount of research measuring LC, few instruments exist that have had their psychometric properties evaluated, and those studies that report statistics tend to report ranges so it is unclear how consistently each feature is counted. The consistency of LC ratings is typically calculated using coefficient α or intraclass correlations (Abedi et al., 2010; Haag et al., 2013; Heppt et al., 2015; Lee & Randall, 2011; Shaftel et al., 2006). Haag et al.'s and Heppt et al.'s studies that examined the count of linguistic features used two-way random effects models to calculate intraclass correlation coefficients and reported the range of intraclass correlation coefficients. Haag et al. reported their coefficients ranged from .79 for counting noun phrases to 1.00 for counting total number of words and Heppt et al. reported their coefficients ranged from .75 for counting academic vocabulary (general and specialized) and 1.00 for counting total number of words, sentences, and words with at least three syllables. Instead of having their raters count or individual features, Lee and Randall (2011) rated items on their lexical and grammatical complexity holistically by having raters rate the items on a scale of one to five. The resulting intraclass correlation coefficients were .31 for lexical complexity ratings and .42 for grammatical complexity ratings. Given the large range in these intraclass correlation coefficients, it is unclear what features can be rated consistently and what features researchers should attend to teaching to their raters to improve the reliability of this instrument.

Although not developed with the intent to measure how item LC affects EBs, Abedi et al. (2010; 2012) developed a rubric for measuring the accessibility of reading assessments for students with disabilities. Specifically, the rubric evaluates the cognitive, grammatical, lexical, and textual/visual features of the items; these dimensions were empirically supported with factor analysis. Part of Abedi et al.'s study examined the reliability of counts of grammatical features with coefficient α; these coefficient alphas ranged from .69 for counting relative clauses to .91 for counting complex verbs. Lexical and grammatical features were adapted from Shaftel et al. (2006) and raters were trained systematically to achieve acceptable reliability using coefficient α. Shaftel et al. (2006) created a linguistic complexity checklist designed for counting the instances particular linguistic features appeared in an assessment item, serving as a holistic measure of LC. Content experts reviewed the checklist familiar with the linguistic features EBs have difficulty with, including mathematics teachers, specialists in mathematics assessments, and an expert in second language learning. While researchers do demonstrate the validity, or accuracy, of their LC measures by utilizing content experts to determine the appropriateness of specific linguistic features used in measuring LC, there is little consistency between studies as to what linguistic features are necessary to include in measuring the effect LC in items on assessment performance, regardless of whether individual features or composites are examined, although counting lexical and grammatical features appears to be fairly reliable for trained raters.

Many of the linguistic features included in Abedi et al.'s (2010) reading accessibility rubric (lexical and grammatical dimensions only) and Shaftel et al.'s (2006) linguistic complexity checklist have been studied in EB assessment research as briefly summarized below.

Lexical Features

Word Frequency and Familiarity

Uncommon words are the most common linguistic feature discussed when determining the extent to which LC influences EB assessment performance. Uncommon words refer to the vocabulary in an item a test-taker may be unfamiliar with; this includes both general academic vocabulary and subject-specific (or technical) vocabulary (Butler et al., 2004). If the uncommon words in an assessment are not content-related, then CIV may be introduced for EBs that are not present for non-EB test-takers (Abedi, 2015). To resolve potential construct-irrelevant issues, test developers are encouraged to use more accessible language and vocabulary that is likely to be understood by all students, such as vocabulary used in a school environment (i.e., pencils and books instead of racquets and badminton). Avenia-Tapper and Llosa (2015), however, caution against modifying the language in items too much, as this may impede measuring construct-relevant vocabulary that is subject-specific.

When measuring uncommon words present on an item, researchers tend to count the instances of general academic vocabulary, although the operationalization of general academic vocabulary varied (Haag et al., 2013; Heppt et al., 2015; Kachchaf et al., 2016; Wolf & Leon, 2009). Butler et al. propose the following protocol for coding academic vocabulary:

- Code phrases and compound words as a single unit
- Distinguish between general academic vocabulary and technical vocabulary
- Infer whether a word with multiple meanings is intended to refer to an academic concept or a common definition
- Distinguish between arbitrary proper names and academic concepts
    - e.g., *Sally* vs. *Alexander Hamilton*

After reaching 80% simple agreement on sample texts, the two coders in Butler et al.'s study began coding the text selections in the study and inter-rater reliability was evaluated with simple agreement. The coders appeared to distinguish between academic and non-academic vocabulary fairy reliably (reliability was above .91 for all subjects evaluated: mathematics, science, and social studies). However, raters appeared to be in less agreement on whether words identified as academic vocabulary were examples of general academic vocabulary or technical vocabulary. Single agreement averaged ".84 (.74-.90) for mathematics, .94 (range .76-1.0) for science, and .91 (range .81-.97) for social studies." The variability in agreement here is concerning for researchers adopting Butler et al.'s protocol as this suggests raters' consistency may vary by subject matter. Some studies reference Butler et al.'s (2004) work on defining the construct of academic vocabulary (Haag et al., 2013; Heppt et al., 2015; Wolf & Leon, 2009).

In Haag et al.'s (2013) and Heppt et al.'s (2015) studies, raters considered the words' definitions and judged whether the students in their targeted population were more likely to encounter the word in a school-based context than elsewhere. If a word was judged to be more likely to be encountered in a school-based context and was not unique to one subject, the word was coded as academic vocabulary. Inter-rater reliability was evaluated with intraclass correlation coefficients. Wolf and Leon (2009) did not disclose in their paper how they classified general academic vocabulary words beyond defining general academic vocabulary as words appearing in multiple subjects. Although Haag et al. (2013) and Heppt et al. (2015) provided intraclass correlation coefficients to provide inter-rater reliability evidence for their study, this method of operationalizing general academic vocabulary is subjective and difficult to replicate. Identifying general academic vocabulary using a corpus such as the Academic Word List (Coxhead, 2000) as Kachchaf et al. (2016) did may provide a less biased estimate on the count of general academic vocabulary.

Martiniello (2008) identified uncommon words in items exhibiting a high amount of DIF, and DIF against EBs is significantly correlated with general academic vocabulary (Haag et al., 2013; Heppt et al., 2015) in some studies, but not in Kachchaf et al.'s, although low-frequency nontechnical vocabulary was found to correlate with DIF (2016). In a study examining differential bundle functioning, on items bundled by the amount of general academic vocabulary, Wolf and Leon (2009) found significant correlations between general academic vocabulary and DIF on bundles containing items with lower item difficulty, but not on bundles containing items with higher item difficulty.

Total number of words

This feature refers to the total number of words contained in an item. The extent to which the total number of words contained in an item influences EB assessment performance is unclear due to inconsistent findings across studies. Wolf and Leon (2009) identified significant correlations with DIF against EBs on bundles containing items with lower item difficulty and Martiniello (2008) found the least LC items had the least number of words. Lee and Randall (2011) found as item length increased, the item was less indicative of math ability for EBs than for non-EBs. In their systematic review of EBs and mathematical word problem solving, Clinton et al. (2018) posited longer items could contain "helpful or irrelevant information," leading to these inconsistent results (p. 192).

Word length

There is limited research on the effect of long words (those with seven or more letters or three or more syllables) on DIF between EBs and non-EBs. Martiniello (2008) compared the most and lost LC items on a content assessment and found the least LC items had shorter words. Heppt et al. (2015) reported significant correlations between the number of words with more than three syllables and DIF against EBs.

Grammatical Features

Passive Voice/Verbs

In sentences with passive voice, the subject receives the action of the verb instead of the subject performing the action (active voice). There are mixed results on whether passive voice influences EB assessment performance. Buono and Jang (2021) found passive voice to be a significant predictor of DIF against EBs and Banks et al. (2016) found differential bundle functioning against EBs in items with passive voice. However, Martiniello (2008) did not identify passive voice in any items flagged for DIF, although she noted passive voice was present in an item ranked low in LC, remarking that the simple sentence structure of the item may have helped EB students' interpretation of the item.

Complex Verbs

Complex verbs have been studied limitedly, although complex verbs were predicted to be influential on the difficulty of an item (Shaftel et al., 2006). Shaftel et al. (2006) defined complex verbs as those "with at least three words ('had been going,' 'would have eaten'), which suggests the use of multiple or difficult verb tenses" (p.121). Abedi et al. (2010) defined complex verbs similar, highlighting that these verbs "are multi-part with a base or main verb and several auxiliaries" (p. 65). Martiniello (2008) identified a complex verb in an item exhibiting high DIF against EBs, yet it appears no study has looked at whether complex verbs are predictive of or correlated to DIF.

Subordinate Clauses

Subordinate clauses, also known as adverbial clauses, are dependent clauses that act as adverbs and begin with a subordinate conjunction (Abedi et al., 2010). These clauses are a more commonly studied linguistic feature, with evidence that suggests subordinate clauses may not influence EB assessment performance. Buono & Jang (2021) did not find subordinate clauses to be a significant predictor of DIF. Items with conditional clauses in Lee and Randall's study did not show evidence of DIF (2011) and subordinate clauses were not correlated with DIF in Kachchaf et al.'s (2016) study. Similarly, Banks et al. (2016) found DBF against EBs in items with conditional clauses, a type of subordinate clause.

Relative Clauses

Relative clauses are a type of subordinate clause that begin with a relative pronoun. These clauses identify and classify nouns or pronouns are also called adjective clauses (Abedi et al., 2010). Like subordinate clauses, there is limited support for relative clauses predicting EB assessment performance. Kachchaf et al. (2016) did not find significant correlations with DIF against EBs and relative clauses, and in Buono & Jang (2021), relative clauses were not a significant predictor of DIF. Banks et al. (2016) also did not find DBF in items with only relative clauses, but the authors suspected a possible cancellation effect may have been masking the effect of relative clauses. However, Loughran (2014) found relative clauses predicted uniform DIF against EBs for fourth graders and relative clauses predicted uniform DIF that favored EBs for eighth graders.

Complex Noun Phrase

Noun phrases consist of a noun and its modifiers and determiners, but are operationalized differently between studies examining the effects of linguistic complexity on EBs. Some studies consider the length of nominals (Buono & Jang, 2021), whereas others count the number of noun phrases (Haag et al., 2013; Kachchaf et al., 2016), with some studies only counting the number of complex noun phrases (Martiniello, 2008). One study found the number of noun phrases predicts DIF against EBs (Haag et al., 2013), but another study found no significant correlations between the number of noun phrases and DIF against EBs (Kachchaf et al. 2016). Although Martiniello (2008) identified complex noun phrases in an item with high levels of DIF against EBs, it is unclear whether noun phrases are significant predictors of DIF. Abedi et al. (2010) defined complex noun phrases as noun phrases with the addition of combinations of determiners, modifiers, and prepositional phrases. Heppt et al. (2015) found a significant correlation between the number of prepositional phrases and DIF against EBs. Martiniello (2008) supports this finding; when conducting think-a-loud protocols in items with high and low levels of DIF, she found some students had difficulty understanding items with prepositional phrases.

Present Study

Although other linguistic features have been studied in EB assessment research, few have been studied as extensively than the features listed above. There appear to be mixed results on whether these features contribute to unnecessary LC in assessment items that unfairly affect EBs over non-EBs. Solano-Flores's (2014) theory on language as a probabilistic phenomenon might explain these mixed findings. Language and language proficiency tend to be viewed in distinct categories (i.e., this assessment item is more linguistically complex than the average item, or this student is an English language learner and therefore is not proficient in English), however this does not acknowledge the different language backgrounds each EB brings with them into a test-taking situation. Many of the studies looking at what specific features contribute to unnecessary LC in assessment items leading to DIF between EBs and non-EBs do not account for how these specific features work together or describe the characteristics of the EBs assessed.

Using a method of measuring LC split into lexical and grammatical complexity via factor analysis may better illustrate the cumulative effects of LC on EB assessment performance, as well as account for the random, or probabilistic, nature of language in assessment items.

However, a comprehensive instrument measuring lexical and grammatical complexity in assessment items has yet to be validated for use in measuring LC that may influence the item responses of EBs. Such an instrument would also need to evaluate whether the lexical and grammatical features identified are construct-relevant, as the presence of more complex lexical and grammatical features is not necessarily irrelevant to the measured construct (Avenia-Tapper & Llosa, 2015). The language that is assessed and is expected to be understood to demonstrate proficiency in the assessed construct is construct-relevant language (Olivieri, 2019). For example, the impact of scientific vocabulary on a biology test may be greater for EBs than for their English-proficient peers, but just because this impact exists does not mean it is a source of item bias or construct-irrelevant variance that is unfairly introduced to the construct. Construct-irrelevant language is language that is not needed to be understood to answer the item correctly, and by evaluating the linguistic complexity of items, construct-relevant and construct-irrelevant language can be disentangled (Abedi, 2015).

Methodology

When it comes to designing or modifying an instrument, the reliability and validity of the measure should be established before use, including the consistency of ratings provided by trained raters. Therefore, the present study utilized generalizability theory, rooted in classical test theory, to determine how many raters are needed to have a reliable and valid measure of grammatical complexity. The measurement of LC is comprised of many factors, and partitioning the measurement error with generalizability theory is a promising way to untangle this source of error (Solano-Flores, 2014; Solano-Flores et al., 2014).

The present study sought to establish reliability evidence for measuring the count of grammatical features expected to influence EB assessment performance by conducting a multivariate generalizability (G) theory decision (D) study to identify different sources of variation in raters, items, and grammatical features. The D study determined the number of raters needed to use the rubric to achieve acceptable reliability. As lexical and grammatical complexity are two differing constructs, separate instruments are needed to measure these two types of linguistic complexity, but a D study was not conducted for measuring lexical complexity as the instrument for measuring lexical features of items (such as uncommon words, total words in an item, total words with seven or more letters) requires less subjective ratings as it was mostly computer-scored and can be determined by one rater, with another rater confirming the first rater's decisions and checking for imputation errors. After completing the D study for grammatical complexity, factor analyses were conducted to confirm the factor structure of the rubric and obtain factor scores for grammatical feature counts and lexical complexity.

Data Collection

The Massachusetts Department of Elementary and Secondary Education (DESE) has a selection of released test items available for the Massachusetts Comprehensive Assessment System (MCAS), an annual statewide assessment administered to students for evaluating school performance. MCAS scores are used to evaluate the performance of and make inferences about EBs, therefore MCAS was an appropriate source of items to evaluate for the effects of unnecessary LC in MCAS items on EB assessment performance. Publicly available student-level item responses were also collected and the effects of LC of MCAS items on student-level item responses were analyzed in Study 2. Four full-length MCAS assessments with fully-released items were selected to rate items for lexical and grammatical complexity; two tenth grade mathematics assessments with 42 items each administered in 2018 and 2019 and two high school biology assessments with 45 items each administered in 2018 and 2019 (DESE, 2019a, 2019b, 2020a, 2020b). Released items from other years of these assessments were used for rater training and practice. Mathematics and science are content areas where construct-irrelevant variance from LC may be introduced for EBs, unlike English language arts assessments where English proficiency is not construct-irrelevant. Assessments administered to high school students were selected because of the characteristics of EBs at this grade level as Study 2 used the LC ratings to evaluate how LC affected EBs differentially based on long-term EB status (five or more years identified as an EB), and first language.

Rater Recruitment and Training

Four raters (including the author) were recruited to score the items for grammatical complexity. These raters were graduate students in education with self-identified native or near-native proficiency in English. Raters were compensated based on the hourly rate graduate students are paid as teaching assistants: $35 an hour for approximately ten hours of participation. Raters were trained how to identify and count five linguistic features identified as contributing to grammatical complexity (passive voice verbs, complex verbs, subordinate clauses, relative clauses, and complex noun phrases). Rater training was completed during a one-hour individual Zoom training with the author; raters reviewed the rater training manual with the author (Appendix A) and completed practice ratings on released items from the 2017 MCAS high school biology assessment. After training was completed, each rater was given a binder with printed copies of the items for scoring. Since the author was one of the raters for the study, before recruitment and training of the raters, the author completed their own counts of grammatical features so they would not be influenced by the counts of other raters.

Coding Grammatical Features

Items were rated using a similar procedure to Abedi et al. (2010); see Appendix A for the training manual on coding grammatical features. For each item, the total number of times a feature is present was recorded. However, because the goal of the overarching dissertation was to measure the effect of construct-irrelevant linguistic complexity on DIF between EBs and non-EBs, the count of the number of times a feature includes construct-relevant vocabulary was also recorded. Construct-relevant vocabulary includes mathematics vocabulary on a mathematics assessment and biology vocabulary on a biology assessment. Construct-relevant vocabulary was obtained by reviewing Massachusetts education standards for the 2017-18 and 2018-19 school years. Raters were provided a word-list of construct-relevant vocabulary for each subject (Appendices B & C). By subtracting the construct-relevant count of a feature from the total count of the same feature, a construct-irrelevant count of a feature can be identified. As construct-relevant language is essential to measuring the focal construct, so are the grammatical features that the construct-relevant language are embedded in. Some features were expected to have more instances of construct-relevant language than others. For example, passive voice and complex verbs tend to be shorter and the language within these grammatical features tends to be verbs while most construct-relevant vocabulary consisted of nouns. On the other hand, subordinate clauses, relative clauses, and particularly complex noun phrases, tend to be lengthier and include nouns, many of which are construct-relevant. Thus, the language in these features needs to be understood to answer the assessment correctly; despite the impact on EBs this language is construct-relevant, not construct-irrelevant (Abedi, 2015; Avenia-Tapper & Llosa, 2015). The counts of construct-irrelevant features were retained for use in the factor analysis for LC. A sample grammatical complexity coding form is shown in Figure 2.1.

Figure 2.1

Sample Grammatical Complexity Coding Form

Grammatical Complexity Code Form										
Rater:										
Subject (circle): Math Biology							**Year (circle):** 2018 2019			
Item #	**Passive (PV) Count**		**Complex Verb (CV) Count**		**Subordinate (SC) Count**		**Relative (RC) Count**		**Noun Phrase (NP) Count**	
	Total	**CR**	**Total**	**CR**	**Total**	**CR**	**Total**	**CR**	**Total**	**CR**
1										
2										
3										

Coding Lexical Features

"Uncommon Words" are the most commonly examined lexical feature in relation to item bias differentially affecting EBs instead of non-EBs. "Uncommon words" refers to words test-takers may be less familiar because these words are used less frequently than other words. Not all uncommon words as construct irrelevant. Rather, some of these uncommon words may be construct relevant; therefore, when identifying uncommon words, words that are construct relevant (such as biology vocabulary on a biology test) should be considered technical vocabulary and words that are construct-irrelevant (such as biology vocabulary on a mathematics test) should be considered general academic vocabulary.

Having raters identify words that are uncommon in assessment items may be particularly unreliable when non-content experts are rating the items. Some researchers instead identify uncommon words by comparing the words in items to a corpus of common words (Abedi et al., 2010; Kachchaf et al., 2016; Wolf & Leon, 2009). In the present study, an online tool to conduct lexical text analysis was used, the Web VocabProfiler Classic (Cobb, 2008; Heatley et al., 2002). This tool recategorizes submitted text across four frequency bands: 1) words from the 1000th most frequent words families, 2) words from the second 1000th most frequent word families, 3) words from the Academic Word List (AWL; Coxhead, 2000), and 4) words that do not appear on the other lists (off-list). As the first two frequency bands identify the most common words, words on the AWL or off-list were considered "uncommon words."

For each item, the number of unique uncommon words were counted and partitioned into unique technical vocabulary word count and unique general academic vocabulary word count. Each assessment item was first submitted to the VocabProfiler tool. Words that were on the AWL or off-list were considered technical vocabulary if the words are biology vocabulary on biology tests and mathematics vocabulary on mathematics tests. Words that were on the AWL or off-list and were not technical vocabulary were considered general academic vocabulary.

The total number of words and total number of words with seven or more letters were also counted. Each item was transcribed to a Microsoft Word document, with equations and expressions (as relevant) removed and replaced with "equation," "expression," etc. in order to treat the unit information as "one word." Microsoft Word's word count feature was used to count the number of words in each item. For each item, each word with seven or more letters in each item was counted once. A sample lexical complexity coding form is shown in Figure 2.2.

Figure 2.2

Sample Lexical Complexity Coding Form

Lexical Complexity Code Form				
Rater:				
Subject (circle): Math Biology		**Year (circle):** 2018 2019		
Item #	**Total Number of Words**	**Unique Technical Vocab Count**	**Unique General Academic Vocab Count**	**7+ Letters Count**
1				
2				
3				

A second rater verified the accuracy of the counts for unique technical vocabulary and unique general academic vocabulary. The second rater reviewed the output of the VocabProfiler tool for each item (specifically words appear on the AWL and off-list frequency bands) and coded the uncommon words. Mismatches in counts for each item for unique technical vocabulary and unique general academic vocabulary were resolved through discussion. The counts of total number of words, unique general academic vocabulary, and words with seven or more letters were retained for use in the SEM for LC.

Data Analysis

Generalizability Theory

Under a classical test theory framework, a test-taker's observed score is the sum of the test-taker's true score and measurement error (Bandalos, 2019). G theory can be used to partition out the sources of measurement error as facets, as measurement error can be influenced by many factors, such as the context of the testing situation or the rater scoring an assessment item (Brennan, 2001). Within generalizability theory, two types of studies can be carried out:

generalizability, or G, studies and decision, or D, studies. G studies look at how measurement error is distributed amongst facets whereas D studies determine how many raters, tests, etc. can be used to reliably estimate a construct.

The present study is a multivariate single-facet D study, with items fully crossed with raters ($i^{\bullet}$ x $R^{\bullet}$) and items and raters crossed with grammatical features (f). Fully crossed facets are desirable because a source of measurement error can be more clearly determined compared to nested facets, which may omit a source of measurement error. For example, if not all raters rated all items, then the interaction between rater and item cannot be examined. Items and raters were treated as random facets as the items and raters in the study were a sample of all possible content assessment items and raters. Linguistic features were treated as fixed facets as these five specific grammatical features were chosen to be measured in the study. When designing a D study, the universe of admissible observations and universe of generalization needs to be taken into consideration. The universe of admissible observations refers to the characteristics of the measured facets and the universe of generalization refers to the characteristics of the universe a researcher attributes their results to (Brennan, 2001). The present study's universe of admissible observations were any content assessment items, raters with a bachelor's or higher who are trained to count the presence of specific grammatical features in assessment items and have self-attested to native or native-like English proficiency, and five grammatical features (passive voice verbs, complex verbs, subordinate clauses, relative clauses, and complex noun phrases). The universe of generalization is the same as the universe of admissible observations as the same sample of items and raters will be used for the G study and the D study.

The present study's design is analogous to a univariate two-facet design (i x R x f), but with a multivariate design, the variance-covariance matrices for each feature can be evaluated.

As each grammatical feature is unique and each grammatical feature count is for a different construct, it would not be appropriate to treat these features as measuring the same construct, which is the count of a particular grammatical feature. Equation 1 lists the equations for all five grammatical features (n_f = 5):

$$\begin{aligned} X_{irf1} &= \mu_{f1} + v_{i1} + v_{r1} + v_{ir1}, \\ X_{irf2} &= \mu_{f2} + v_{i2} + v_{r2} + v_{ir2}, \\ X_{irf3} &= \mu_{f3} + v_{i3} + v_{r3} + v_{ir3}, \\ X_{irf4} &= \mu_{f4} + v_{i4} + v_{r4} + v_{ir4}, \\ X_{irf5} &= \mu_{f5} + v_{i5} + v_{r5} + v_{ir5}. \end{aligned} \tag{1}$$

In these equations, X_{ir} represents the observed count of a grammatical feature for a given item for a given rater and μ represents the grand mean for a feature across all items and raters. v_i represents an item's effect, v_r represents a rater's effect, and v_{ir} represents the interaction effect between item and rater. The variability in the observed counts of grammatical features, or observed score variance of a feature, can be partitioned as follows in Equation 2.

$$\begin{aligned} \sigma^2_{irf1} &= \sigma^2_{i1} + \sigma^2_{r1} + \sigma^2_{ir1} + \sigma^2_{ir,e1}, \\ \sigma^2_{irf2} &= \sigma^2_{i2} + \sigma^2_{r2} + \sigma^2_{ir2} + \sigma^2_{ir,e2}, \\ \sigma^2_{irf3} &= \sigma^2_{i3} + \sigma^2_{r3} + \sigma^2_{ir3} + \sigma^2_{ir,e3}, \\ \sigma^2_{irf4} &= \sigma^2_{i4} + \sigma^2_{r4} + \sigma^2_{ir4} + \sigma^2_{ir,e4}, \\ \sigma^2_{irf5} &= \sigma^2_{i5} + \sigma^2_{r5} + \sigma^2_{ir5} + \sigma^2_{ir,e5}. \end{aligned} \tag{2}$$

The variance and covariance components for the universe of admissible observations is a summation of three unstructured covariance matrices, Σ_i, Σ_r, and Σ_{ir}, as listed below.

$$\Sigma_i = \begin{bmatrix} \sigma^2_{1i} & \sigma^2_{1i2i} & \sigma^2_{1i3i} & \sigma^2_{1i4i} & \sigma^2_{1i5i} \\ \sigma^2_{1i2i} & \sigma^2_{2i} & \sigma^2_{2i3i} & \sigma^2_{2i4i} & \sigma^2_{2i5i} \\ \sigma^2_{1i3i} & \sigma^2_{2i3i} & \sigma^2_{3i} & \sigma^2_{3i4i} & \sigma^2_{3i5i} \\ \sigma^2_{1i4i} & \sigma^2_{2i4i} & \sigma^2_{3i4i} & \sigma^2_{4i} & \sigma^2_{4i5i} \\ \sigma^2_{1i5i} & \sigma^2_{2i5i} & \sigma^2_{3i5i} & \sigma^2_{4i5i} & \sigma^2_{5i} \end{bmatrix}$$

$$\Sigma_r = \begin{bmatrix} \sigma^2_{1r} & \sigma^2_{1r2r} & \sigma^2_{1r3r} & \sigma^2_{1r4r} & \sigma^2_{1r5r} \\ \sigma^2_{1r2r} & \sigma^2_{2r} & \sigma^2_{2r3r} & \sigma^2_{2r4r} & \sigma^2_{2r5r} \\ \sigma^2_{1r3r} & \sigma^2_{2r3r} & \sigma^2_{3r} & \sigma^2_{3r4r} & \sigma^2_{3r5r} \\ \sigma^2_{1r4r} & \sigma^2_{2r4r} & \sigma^2_{3r4r} & \sigma^2_{4r} & \sigma^2_{4r5r} \\ \sigma^2_{1r5r} & \sigma^2_{2r5r} & \sigma^2_{3r5r} & \sigma^2_{4r5r} & \sigma^2_{5r} \end{bmatrix}$$

$$\Sigma_{ir} = \begin{bmatrix} \sigma^2_{1e} & \sigma^2_{1e2e} & \sigma^2_{1e3e} & \sigma^2_{1e4e} & \sigma^2_{1e5e} \\ \sigma^2_{1e2e} & \sigma^2_{2e} & \sigma^2_{2e3e} & \sigma^2_{2e4e} & \sigma^2_{2e5e} \\ \sigma^2_{1e3e} & \sigma^2_{2e3e} & \sigma^2_{3e} & \sigma^2_{3e4e} & \sigma^2_{3e5e} \\ \sigma^2_{1e4e} & \sigma^2_{2e4e} & \sigma^2_{3e4e} & \sigma^2_{4e} & \sigma^2_{4e5e} \\ \sigma^2_{1e5e} & \sigma^2_{2e5e} & \sigma^2_{3e5e} & \sigma^2_{4e5e} & \sigma^2_{5e} \end{bmatrix}$$

These variance and covariance components were estimated in mGENOVA (version 2.1), a program designed to run multivariate G and D study analyses (Brennan, 2001). Elements from the variance and covariance components matrices can be used to calculate generalizability coefficients and dependability coefficients for each grammatical feature. Generalizability coefficients represent a norm-referenced coefficient of reliability and dependability coefficients represent a criterion-referenced coefficient of reliability. The specific contributions of each feature to a grammatical complexity composite score were determined through factor analysis, which is described in the next section. The calculations of these coefficients are like a univariate D study due to the balanced design of the present study. In the present study, for a grammatical feature f, the generalizability coefficient ρ^2_f is calculated as shown in Equation 3 and the dependability coefficient ϕ_f is calculated as shown in Equation 4. To determine the number of raters needed for acceptably reliable estimation of the counts of grammatical features, the variance in observed counts attributable to raters needs to be adjusted, with n'_r denoting the adjustment to the number of raters. Acceptably reliable coefficients are at or above .800, although higher cut-offs are encouraged if decisions will have greater consequences (Webb et al., 2007).

$$\rho_f^2 = \frac{\sigma_{1i}^2}{\sigma_{1i}^2 + \frac{\sigma_{1e}^2}{n'_r}} \tag{3}$$

$$\phi_f = \frac{\sigma_{1i}^2}{\sigma_{1i}^2 + \frac{\sigma_{1r}^2}{n'_r} + \frac{\sigma_{1e}^2}{n'_r}} \tag{4}$$

Determining Composites of Linguistic Complexity

After determining the reliability of counting lexical and grammatical features, these construct-irrelevant counts were standardized (with each feature's count transformed into z-scores) for use in confirmatory factor analyses (CFAs). Factor scores extracted from these CFAs can be used to create composite scores for lexical complexity and grammatical complexity; a composite score of LC comprising of lexical and grammatical factors was also explored. The standardized factor scores from these factor analyses can be used as a measure of lexical or grammatical complexity in an item (Shavelson et al., 1989). While these individual features can be used as predictors, establishing composite scores of LC can allow researchers to predict the cumulative effects of lexical and grammatical complexity. In the present study, each rater's count of each item is used to determine composite scores of LC; each rater's count of a feature is its own variable.

Steps need to be followed to determine if it is appropriate to use the data to model lexical and grammatical complexity as a multidimensional or higher-order CFA. Credé and Harms (2015) argue more parsimonious alternative models should be examined before settling with a multidimensional or higher-order model structure. First, a unidimensional model with all observed indicators (counts of lexical and grammatical features) loading onto one factor for LC was tested for each subject. This model's fit statistics were then compared to those of a corresponding six-dimensional model with all observed indicators loading onto their specific features' factors (e.g., lexical features load onto a lexical complexity factor, passive voice counts for each rater load onto a passive voice factor, complex verb counts for each rater load onto a complex verb factor, etc.). If the six-dimensional model is better fitting than the unidimensional model, then LC is multidimensional. Due to problems with consistency in raters' counts of some features (discussed in the results section), other unidimensional and multidimensional models omitting those features were explored for both subjects.

After determining multidimensionality and the number of latent constructs, measurement models for each factor in each subject were conducted to evaluate the fit of each factor. This led to the creation of new multidimensional models as some measurement models were just-identified. Fit cannot be determined without an over-identified model, so more parameters were fixed to evaluate model fit (Kline, 2023). Next, I tested whether my models a higher-order model fit the data better than a multidimensional model. A higher-order model is a CFA where factors may load onto higher-order factors. To establish a composite of LC, a model with the passive voice, complex verb, subordinate clause, relative clause, complex noun phrase, and lexical complexity factors loading onto an LC factor must fit better than a multidimensional model. To establish a composite of grammatical complexity, a model with the passive voice, complex verb, subordinate clause, relative clause, and complex noun phrase factors loading onto a grammatical complexity factor must fit better than a multidimensional model (this multidimensional model would not include lexical complexity). The best-fitting, most parsimonious model should be selected.

If the higher-order model for LC fits the data the best, then LC factor scores would be used in further analyses in Study 2. If the higher-order model for grammatical complexity fits best, then grammatical complexity factor scores would be used in further analyses; lexical complexity factor scores from the measurement model would be used in further analyses. If the multidimensional model fits best, factor scores from each measurement model included in the multidimensional model would be used in further analyses. Final models were selected based on cutoff criteria for fit indices from Schreiber et al. (2006), who reviewed and provided guidelines for fit criteria. According to Schrieber et al., well-fitting models have RMSEA $\leq$ .08 (with reported confidence interval), CFI $\geq$ .95, and SRMR $\leq$ .08.

Results

In this section, the results of the decision study and CFAs will be presented for the biology and mathematics assessments. The total counts and construct-irrelevant counts of features are presented side-by-side for a subject as the present study seeks to determine if features are determined to be construct-irrelevant reliably across raters. If construct-irrelevant counts are less reliable than total counts, raters may not be identifying construct-relevant vocabulary accurately. The first results shown are the mean grammatical feature counts per item by rater and subject, followed by a discussion on the whether the raters in the study were adept at identifying the grammatical features, and the generalizability and dependability coefficients for the raters participating in the study. Next, the variance and covariance components for the biology and mathematics assessments’ feature counts are discussed. The last results from the decision study concern the number of raters required to reliability count grammatical features on these assessments based on estimated generalizability and dependability coefficients are presented, along with the variance components from the decision study.

After the results of the decision study are descriptive statistics for the coding of lexical features on the assessments. This is followed by the LC CFAs, where two-factor and one-factor model results are presented and compared. Although LC is theorized to have multiple factors (in the present study, LC is partitioned into lexical and grammatical complexity), due to the high correlation between factors in the two-factor models, I examined one-factor models of LC where lexical and grammatical features all loaded onto the same factor.

Coding Grammatical Features and Decision Study

Feature Counts and Initial Generalizability and Dependability Coefficients

Table 2.1 presents the mean count per item for all five grammar features for each subject, averaged across raters. Mean counts of feature before accounting for construct-relevant or irrelevant vocabulary appeared low or near zero, suggesting some raters may have been under-identifying features. Table 2.2 presents the number of items raters identified a feature in on the Biology and Mathematics assessments. The number of items with a particular feature should be constant across raters if features are coded correctly, however if raters are not adept at recognizing features, they will have a lower count that other raters. It was determined a rater's count of a feature would be "too low" if a rater's number of items with a given feature was less than one standard deviation below the mean of all four raters. Alternatively, if a rater has a higher count than other raters, they may be over-identifying these features. It was decided a rater's count of a feature would be "too high" if a rater's number of items with a given feature was more than one standard deviation below the mean of all four raters.

Table 2.1.

Mean Grammatical Feature Counts Per Item by Rater and Subject

Feature	Test – Total/CIR	R1	R3	R4	R6	Grand Mean
PV	Biology – Total	0.200	0.789	0.789	0.556	0.583
	Biology – CIR	0.200	0.767	0.322	0.533	0.456
	Math – Total	0.095	0.631	0.476	0.393	0.399
	Math – CIR	0.083	0.548	0.155	0.321	0.277
CV	Biology – Total	0.111	0.344	0.167	0.222	0.211
	Biology – CIR	0.111	0.333	0.156	0.211	0.203
	Math – Total	0.012	0.119	0.083	0.250	0.116
	Math – CIR	0.012	0.119	0.060	0.238	0.107
SC	Biology – Total	0.622	1.011	0.033	1.056	0.681
	Biology – CIR	0.167	0.389	0.022	0.456	0.258
	Math – Total	0.131	0.464	0.167	0.357	0.280
	Math – CIR	0.083	0.440	0.131	0.119	0.193
RC	Biology – Total	0.789	0.933	0.100	0.878	0.675
	Biology – CIR	0.411	0.622	0.089	0.433	0.389
	Math – Total	0.512	0.690	0.012	0.679	0.473
	Math – CIR	0.393	0.524	0.000	0.429	0.336
NP	Biology – Total	4.722	6.122	2.100	3.800	4.186
	Biology – CIR	1.733	2.611	1.467	1.189	1.750
	Math – Total	4.250	4.333	2.131	3.369	3.521
	Math – CIR	1.524	1.833	1.012	0.976	1.336

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase. “Total” refers to the total feature counts averaged across items and “CIR” refers to the construct-irrelevant features counts averaged across items.

Table 2.2.

Number of Items in a Subject Identified with a Given Feature by Rater.

Feature	Subject	R1	R3	R4	R6	1 SD below mean	1 SD above mean
PV	Biology	**12**	38	39	30	17.3	42.3
	Math	**7**	35	29	27	12.3	36.7
CV	Biology	**7**	**27**	14	19	8.3	25.2
	Math	**1**	9	7	**15**	2.2	13.8
SC	Biology	34	44	**3**	52	11.8	54.7
	Math	10	13	13	**21**	9.5	19.0
RC	Biology	39	45	**8**	44	16.5	51.5
	Math	27	33	**1**	36	8.3	40.2
NP	Biology	87	90	**71**	86	75.0	92.0
	Math	83	84	**63**	79	67.5	87.0

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase. Bold numbers denote raters that under-identified or over-identified a feature.

Based on these cut-values, Rater 1 under-identified passive voice and complex verbs for all assessments, Rater 3 over-identified subordinate clauses for the biology assessments, Rater 4 under-identified subordinate clauses for the biology assessments, relative clauses for all assessments, and complex noun phrases for all assessments, and Rater 6 over-identified complex verbs and subordinate clauses for the mathematics assessments. Across all features, Raters 1 and 4 had lower counts than Raters 3 and 6 and appeared to make systematic errors identifying specific features, in many cases not identifying them at all. While Raters 3 and 6 over-identified complex verbs in the biology and mathematics assessments, respectively, according to the cut-values, reviewing their coding suggested they identified complex verbs correctly and they likely identified complex verbs the other raters missed.

Generalizability and dependability coefficients for the rater counts of grammatical features are presented in Table 2.3, assuming four raters and ninety items for the biology assessments and 84 items for the mathematics assessments. As discussed in previously, coefficients were seen as acceptably reliable at .800 (Webb et al., 2007). For the biology assessments, raters were consistent with their total counts of grammatical features, however when it came to deciding whether these features included construct-relevant vocabulary, counts were less reliable. For the mathematics assessments, raters were fairly reliable for the total count of passive voice, relative clauses, and noun phrases. Although relative clause and noun phrase counts with construct-irrelevant vocabulary were reliable; the passive voice counts with construct-relevant vocabulary was much less reliable. The generalizability and dependability coefficients for the complex verb and subordinate clause counts were unreliable, both for the total count and counts with construct-irrelevant vocabulary.

Table 2.3.

Generalizability Coefficients for the Rater Counts of Grammatical Features

Test – Total/CIR	**Generalizability Coefficient**	**PV**	**CV**	**SC**	**RC**	**NP**
Biology – Total	ρ_f^2	0.819	0.762	0.719	0.788	0.862
	ϕ_f	0.801	0.750	0.667	0.746	0.757
Biology – CIR	ρ_f^2	0.712	0.753	0.620	0.711	0.791
	ϕ_f	0.691	0.741	0.595	0.689	0.768
Math – Total	ρ_f^2	0.750	0.432	0.472	0.830	0.936
	ϕ_f	0.725	0.417	0.463	0.792	0.905
Math – CIR	ρ_f^2	0.450	0.366	0.028	0.825	0.849
	ϕ_f	0.423	0.353	0.026	0.804	0.838

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase. “Total” refers to the total feature counts averaged across items and “CIR” refers to the construct-irrelevant features counts averaged across items.

Variance and Covariance Components for Biology Assessments' Feature Counts

Table 2.4 presents the variance and covariance matrices for each component used in generalizability and dependability coefficient calculation for the biology assessments, assuming four raters and 90 items. Percentages show the variation in a feature attributable to that feature's variance component. Readers should note the item by rater interaction includes variation attributable to error.

Table 2.4.

Estimates of Variance and Covariance Components used in Generalizability Coefficient Calculations, Biology Assessments

		Biology - Total					**Biology - CIR**				
		PV	**CV**	**SC**	**RC**	**NP**	**PV**	**CV**	**SC**	**RC**	**NP**
Σ_i	PV	.640	**.503**	**.049**	**.410**	**.083**	.325	**.524**	**.004**	**.287**	**.101**
	CV	.128	.102	**.199**	**.427**	**.223**	.090	.092	**.244**	**.374**	**.296**
	SC	.028	.044	.489	**.200**	**.246**	.001	.027	.134	**.451**	**.657**
	RC	.232	.096	.099	.499	**.428**	.082	.057	.083	.251	**.548**
	NP	.139	.149	.360	.632	4.377	.087	.137	.365	.417	2.310
		50%	43%	33%	42%	44%	36%	42%	27%	36%	45%
Σ_r	PV	.071					.056				
	CV	.019	.009				.024	.008			
	SC	-.017	.027	.216			.037	.013	.036		
	RC	-.038	.017	.176	.145		.036	.014	.035	.044	
	NP	-.075	.098	.605	.560	2.815	.088	.041	.022	.086	.352
		6%	4%	15%	12%	28%	6%	4%	7%	6%	7%
Σ_{ir}	PV	.564					.525				
	CV	-.051	.127				-.028	.121			
	SC	-.025	.023	.763			.007	.006	.328		
	RC	-.003	-.003	-.042	.536		.070	-.007	.047	.407	
	NP	-.063	-.010	-.009	.057	2.792	.042	-.054	.095	.110	2.435
		44%	54%	52%	46%	28%	58%	55%	66%	58%	48%

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase. "Total" refers to the total feature counts averaged across items and "CIR" refers to the construct-irrelevant features counts averaged across items.

For passive voice, the largest sources of variation in the total count on the biology assessments were due to items (50.2% of count variation) and the item by rater interaction (44.2%), with some variation due to raters (5.6%). When considering the construct-irrelevant counts of passive voice, there was decreased variation due to items (35.8%), with increased variation due to raters (6.2%) and the item by rater interaction (58.0%). For complex verbs, the largest sources of variation in the total count on the biology assessments were due to items (42.9% of count variation) and the item by rater interaction (53.5%), with little variation due to raters (3.6%). The distribution of sources of variation for the construct-irrelevant counts of complex verbs was similar to the total count, but with increased variation due to items (41.6%) and the item by rater interaction (54.8%), with decreased variation due to raters (3.6%). Due to the similarity in variance components for total count of features and construct-irrelevant count of features in the biology assessments, there was little variation in counts due to raters for complex verbs. However, this is due to most of variation in counting complex verbs coming from item or item by rater variance components, suggesting there is some feature influencing raters' counts that was not captured, especially as the item by rater variance components (which are large), includes variance associated with error. The variance in counting complex verbs (both total and construct-irrelevant) attributable to items was small compared to the variance components for other features' counts.

For subordinate clauses, the largest sources of variation in the total count on the biology assessments were due to the item by rater interaction (52.0% of count variation) and items (33.3%), with some variation due to raters (14.7%). However, when construct-irrelevant counts of subordinate clauses were considered, there was decreased variation due to items (26.9%) and raters (7.3%), with increased variation due to the item by rater interaction (65.8%). For relative clauses, the largest sources of variation in the total count on the biology assessments were due to the item by rater interaction (45.5% of count variation) and items (42.3%), with some variation due to raters (12.3%). Like subordinate clauses, when construct-irrelevant counts of relative clauses were considered, there was decreased variation due to items (35.7%) and raters (6.3%), with increased variation due to the item by rater interaction (58.0%).

For the total count of noun phrases, most of the variation was attributed to items and raters, most of the variation was captured by the facets examined. For noun phrases, the largest source of variation in the total count on the biology assessments was due to items (43.8%), although the variation attributable to raters (28.2%) and the item by rater interaction (28.0%) was also large. The distribution of sources of variation for the construct-irrelevant counts of noun phrases changed in a way similar to the counts of the other grammatical features. There was decreased variation due to items (45.3%) and increased variation due to raters (6.9%) and the item by rater interaction (47.8%). Generally, the variation in counts attributable to the item by rater interaction was much larger than the variation attributable to items or raters, which suggests there is some aspect of the variation items that influences the raters construct-irrelevant counts that was not captured. Readers might note the variance components for complex noun phrases is large, regardless of variation source (except for variation in construct-irrelevant counts attributable to raters). Raters tended to count more noun phrases than other grammatical features in assessment items, leading to this difference in scale. Features were weighted equally when calculating generalizability coefficients.

The covariance components in Table 2.4 can be examined to look at the relationship between counts of different grammatical features. For variation attributable to items, the covariance attributable to items had some interesting patterns. The total counts and construct-irrelevant counts of subordinate clauses tended to have small covariances with passive voice, complex verbs, and relative clauses. The total counts of complex noun phrases tended to have larger covariances with other grammatical features, although the covariance between the construct-irrelevant counts of passive voice and complex noun phrases was small.

For variation attributable to raters, the covariance attributable to raters tended to be small for some features, and larger for other features. Total and construct-irrelevant counts of passive voice and complex verbs tended to have low covariance with other features. Total counts of passive voice had negative covariance (indicating a negative relationship) with subordinate clauses, relative clauses, and complex noun phrases, although these values are small. The total counts of subordinate clauses, subordinate clauses, and complex noun phrases have larger covariance attributable to raters, but these covariances decrease when looking at the construct-irrelevant counts of these features.

For variation attributable to the interaction between items and raters (or measurement error), there was little covariation among features, although there were negative covariance components for the total counts of passive voice, complex verbs, and subordinate clauses and for the construct-irrelevant counts of passive voice and complex verbs. These negative covariance components indicate a negative relationship between these features, but the covariance is small.

Variance and Covariance Components for Mathematics Assessments' Feature Counts

Table 2.5 presents the variance and covariance matrices for the mathematics assessments, assuming four raters and 84 items. For the mathematics assessments with the counts of construct-irrelevant features, readers may note the correlations for subordinate clauses are greater than one. This is due to how small the subordinate clause variance component for items is (.003), which also lead to low generalizability coefficients for counting construct-irrelevant subordinate clauses (ρ_f^2 = .028 and ϕ_f = .026).

Table 2.5.

Estimates of Variance and Covariance Components used in Generalizability Coefficient Calculations, Mathematics Assessments

		Mathematics - Total					**Mathematics - CIR**				
		PV	**CV**	**SC**	**RC**	**NP**	**PV**	**CV**	**SC**	**RC**	**NP**
Σ_i	PV	.255	**.211**	**.743**	**.321**	**.313**	.070	**.495**	**3.265**	**.341**	**.372**
	CV	.017	.025	**.603**	**.478**	**.772**	.018	.019	**2.271**	**.471**	**.811**
	SC	.126	.032	.112	**.727**	**.736**	.049	.018	.003	**3.178**	**2.425**
	RC	.105	.049	.158	.421	.627	.056	.040	.111	.381	**.639**
	NP	.419	.323	.651	1.077	7.011	.154	.175	.216	.618	2.455
		40%	15%	18%	49%	70%	16%	12%	1%	51%	56%
Σ_r	PV	.047					.038				
	CV	.010	.008				.011	.008			
	SC	.026	.010	.019			.030	.002	.022		
	RC	.001	.013	.033	.097		.029	.008	.017	.049	
	NP	-.046	-.015	.062	.270	1.026	.040	-.014	.049	.059	.152
		7%	5%	3%	11%	10%	9%	5%	5%	7%	4%
Σ_{ir}	PV	.340					.341				
	CV	-.008	.131				-.008	.131			
	SC	.022	.021	.501			.033	.017	.453		
	RC	-.011	.033	.134	.344		-.033	.023	.071	.323	
	NP	-.024	-.080	.057	.094	1.921	.058	-.056	.084	.075	1.748
		53%	80%	79%	40%	19%	76%	83%	95%	43%	40%

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase. "Total" refers to the total feature counts averaged across items and "CIR" refers to the construct-irrelevant features counts averaged across items.

For passive voice, the largest sources of variation in the total count on the mathematics assessments were due to items (39.8% of count variation) and the item by rater interaction (53.0%), with some variation due to raters (7.3%). When considering the construct-irrelevant counts of passive voice, there was decreased variation due to items (15.5%), with increased variation due to raters (8.6%) and the item by rater interaction (75.9%). For complex verbs, the largest sources of variation in the total count on the mathematics assessments were due to the item by rater interaction (79.7% of count variation), with some variation due to items (15.2%) and raters (5.1%). The distribution of sources of variation for the construct-irrelevant counts of complex verbs was similar to the total count, but with decreased variation due to items (12.0%), no change in variation due to raters (5.1%), and increased variation due to the item by rater interaction (83.0%). Due to the similarity in variance components for total count of features and construct-irrelevant count of features in the mathematics assessments, there was little variation in counts due to items for complex verbs. This is due to most of the variation in counting complex verbs coming from the item by rater variance component, suggesting raters were inconsistent overall in their counts of complex verbs on the mathematics assessments.

For subordinate clauses, the largest sources of variation in the total count on the mathematics assessments were due the item by rater interaction (79.3% of count variation), with some variation due to items (17.7%) and raters (3.0%). The distribution of sources of variation for the construct-irrelevant counts of subordinate clauses led to increased variation due to raters (4.6%) and the item by rater interaction (94.7%), with decreased variation due to items (.7%). Most of the variation in counts of passive voice, complex verbs, and subordinate clauses was overwhelmingly due to the item by rater interactions. However, for relative clauses, the largest sources of variation in the total count on the mathematics assessments were due to items (48.8% of count variation) and the item by rater interaction (39.9%), with some variation due to raters (11.3%). The distribution of sources of variation for the construct-irrelevant counts of relative clauses led to increased variation due to items (50.6%) and the item by rater interaction (42.8%), with decreased variation due to raters (6.6%).

For noun phrases, more variation was attributed to items, suggesting raters were more consistent in their counts of noun phrases. For noun phrases, the largest source of variation in the total count on the biology assessments was due to items (70.4%), with some variation attributable to raters (10.3%) and the item by rater interaction (19.3%). The distribution of sources of variation for the construct-irrelevant counts of noun phrases changed in a way like the counts of the other grammatical features. There was decreased variation due to items (56.4%) and increased variation due to raters (3.5%) and the item by rater interaction (40.1%). Generally, the variation in counts attributable to the item by rater interaction was much larger than the variation attributable to items or raters, which suggests there is some aspect of the items that influences the raters counts that was not captured.

The covariance components in Table 2.5 can be examined to look at the relationship between counts of different grammatical features. Readers should note the large correlations between subordinate clauses and other grammatical features, for both the total and construct-irrelevant counts. This is due to the small amount of variation in subordinate clause counts. Similarly, complex verbs had a low amount of variation in counts leading to inflated positive relationships in the counts of other variables. For variation attributable to items, the total counts of complex noun phrases tended to have larger covariances with other grammatical features.

For variation attributable to raters, the covariance attributable to raters tended to be small for most features. Total counts of complex noun phrases had negative covariances with passive voice and complex verbs, and construct-irrelevant counts of complex noun phrases had a negative covariance with complex verbs, although these values are small.

For variation attributable to the interaction between items and raters (or measurement error), the covariance attributable to the interaction between items and raters was small for some features and larger for others. The total and construct-irrelevant counts of passive voice and complex verbs tended to have small covariances with other features. Some of these covariance components were negative, indicating a negative relationship between these features, although the covariance is small.

Number of Raters Required to Reliably Count Grammatical Features

Generalizability and dependability coefficients for the numbers of raters required to reliably score grammatical complexity in assessment items are in Tables 2.6-2.9. The coefficients for four raters are identical to those in Table 2.3, but serve as reference points. Overall, generalizability coefficients (ρ_f^2) were more reliable than the calculated dependability coefficients (ϕ_f) and at times the generalizability coefficients were sufficiently reliable for a lower number of raters than the dependability coefficients. Due to how difficult the task was for these raters (as they were not linguistic content experts) and the purpose of determining LC (identifying norm-referenced scores of LC to identify which items are more linguistically complex than others on an assessment), the generalizability coefficients were used to determine the cut-off for the number of raters needed to consistently count grammatical features. As mentioned previously raters will be considered to reliably count a feature when the generalizability coefficients for a particular number of raters is at or above .800 (Webb et al., 2007). If the purpose of using the Grammatical Complexity Coding Form is to make absolute decisions about the count of grammatical features in items, more raters should be used (if using raters than are not linguistic content experts) or linguistic content experts should be recruited as raters.

Raters were fairly consistent in the total counts of grammatical feature for the biology assessments (Table 2.6), assuming 90 items (the number of items on two biology assessments).

Based on generalizability coefficients, four raters are needed to consistently count passive voice, five raters to consistently count complex verbs, and five raters to consistently count relative clauses and noun phrases. It appears not even six raters were enough to count subordinate clauses consistently. If reliant on absolute decisions about the total counts of grammatical features with these raters, about one more rater would be needed, although subordinate clauses would not be rated consistently, suggesting more training was needed with these raters to reliably count subordinate clauses.

Table 2.6.

Decision Study Variance Components and Generalizability Coefficients for the Rater Counts of Grammatical Features: All Grammatical Features on MCAS Biology Assessments

Feature	Component	Number of raters				
		2	3	4	5	6
PV	σ_i^2	.640	.640	.640	.640	.640
	σ_r^2	.036	.024	.018	.014	.012
	σ_{ir}^2	.282	.188	.141	.113	.094
	ρ_f^2	.694	.773	.819	.850	.872
	ϕ_f	.668	.751	.801	.834	.858
CV	σ_i^2	.102	.102	.102	.102	.102
	σ_r^2	.004	.003	.002	.002	.001
	σ_{ir}^2	.063	.042	.032	.025	.021
	ρ_f^2	.616	.707	.762	.801	.828
	ϕ_f	.601	.693	.750	.790	.819
SC	σ_i^2	.489	.489	.489	.489	.489
	σ_r^2	.108	.072	.054	.043	.036
	σ_{ir}^2	.382	.254	.191	.153	.127
	ρ_f^2	.562	.658	.719	.762	.794
	ϕ_f	.500	.600	.667	.714	.750
RC	σ_i^2	.499	.499	.499	.499	.499
	σ_r^2	.072	.048	.036	.029	.024
	σ_{ir}^2	.268	.179	.134	.107	.089
	ρ_f^2	.650	.736	.788	.823	.848
	ϕ_f	.594	.687	.746	.786	.815
NP	σ_i^2	4.377	4.377	4.377	4.377	4.377
	σ_r^2	1.407	.938	.704	.563	.469
	σ_{ir}^2	1.396	.931	.698	.558	.465
	ρ_f^2	.758	.825	.862	.887	.904
	ϕ_f	.610	.701	.757	.796	.824

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase.

Raters were considerably less consistent in the construct-irrelevant counts of grammatical feature for the biology assessments (Table 2.7), assuming 84 items (the number of items on two mathematics assessments). Based on generalizability coefficients, six raters are needed to consistently count complex verbs and five raters to consistently count noun phrases; the same conclusions may be made from the dependability coefficients. It appears not even six raters were enough to count passive voice, subordinate clauses, and relative clauses consistently. Coefficient estimates for construct-irrelevant counts of complex verbs are similar to those for total counts of complex verbs for the biology assessments, but much less than for those of other features, suggesting raters were not consistently identifying construct-irrelevant vocabulary.

Table 2.7.

Decision Study Variance Components and Generalizability Coefficients for the Rater Counts of Grammatical Features: Construct-Irrelevant Grammatical Features on MCAS Biology Assessments

Feature	Component	Number of raters				
		2	3	4	5	6
PV	σ_i^2	.325	.325	.325	.325	.325
	σ_r^2	.028	.019	.014	.011	.009
	σ_{ir}^2	.263	.175	.131	.105	.088
	ρ_f^2	.553	.650	.712	.755	.788
	ϕ_f	.527	.626	.691	.736	.770
CV	σ_i^2	.092	.092	.092	.092	.092
	σ_r^2	.004	.002	.002	.002	.001
	σ_{ir}^2	.060	.040	.030	.024	.020
	ρ_f^2	.603	.695	.753	.792	.820
	ϕ_f	.588	.682	.741	.781	.811
SC	σ_i^2	.134	.134	.134	.134	.134
	σ_r^2	.018	.012	.009	.007	.006
	σ_{ir}^2	.164	.109	.082	.066	.055
	ρ_f^2	.449	.550	.620	.671	.710
	ϕ_f	.423	.524	.595	.647	.688
RC	σ_i^2	.251	.251	.251	.251	.251
	σ_r^2	.022	.015	.011	.009	.007
	σ_{ir}^2	.204	.136	.102	.081	.068
	ρ_f^2	.552	.649	.711	.755	.787
	ϕ_f	.526	.625	.689	.735	.769
NP	σ_i^2	2.310	2.310	2.310	2.310	2.310
	σ_r^2	.176	.117	.088	.070	.059
	σ_{ir}^2	1.218	.812	.609	.487	.406
	ρ_f^2	.655	.740	.791	.826	.851
	ϕ_f	.624	.713	.768	.806	.833

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase.

Raters were consistent in the total counts of some grammatical feature for the mathematics assessments, but not for other features (Table 2.8). Based on generalizability coefficients, six raters are needed to consistently count passive voice, four raters to consistently count relative clauses, and two raters to consistently count and noun phrases. It appears not even six raters were enough to count complex verbs and subordinate clauses consistently. If reliant on absolute decisions about the total counts of grammatical features with these raters, the same number of raters per feature would be appropriate (even if not exactly meeting the .800 threshold), although complex verbs and subordinate clauses would not be rated consistently, suggesting more training was needed with these raters to reliably count complex verbs and subordinate clauses.

Table 2.8.

Decision Study Variance Components and Generalizability Coefficients for the Rater Counts of Grammatical Features: All Grammatical Features on MCAS Mathematics Assessments

Feature	Component	Number of raters				
		2	3	4	5	6
PV	σ_i^2	.255	.255	.255	.255	.255
	σ_r^2	.023	.016	.012	.009	.008
	σ_{ir}^2	.170	.113	.085	.068	.057
	ρ_f^2	.600	.693	.750	.790	.818
	ϕ_f	.569	.665	.725	.768	.798
CV	σ_i^2	.025	.025	.025	.025	.025
	σ_r^2	.004	.003	.002	.002	.001
	σ_{ir}^2	.066	.044	.033	.026	.022
	ρ_f^2	.276	.364	.432	.488	.533
	ϕ_f	.264	.349	.417	.472	.518
SC	σ_i^2	.112	.112	.112	.112	.112
	σ_r^2	.010	.006	.005	.004	.003
	σ_{ir}^2	.250	.167	.125	.100	.083
	ρ_f^2	.309	.401	.472	.528	.573
	ϕ_f	.301	.392	.463	.518	.564
RC	σ_i^2	.421	.421	.421	.421	.421
	σ_r^2	.049	.032	.024	.019	.016
	σ_{ir}^2	.172	.115	.086	.069	.057
	ρ_f^2	.710	.786	.830	.860	.880
	ϕ_f	.656	.741	.792	.827	.851
NP	σ_i^2	7.010	7.011	7.011	7.011	7.011
	σ_r^2	.513	.342	.257	.205	.171
	σ_{ir}^2	.961	.640	.480	.384	.320
	ρ_f^2	.879	.916	.936	.948	.956
	ϕ_f	.826	.877	.905	.922	.935

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase.

Raters were considerably less consistent in the construct-irrelevant counts of grammatical feature for the mathematics assessments (Table 2.9). Based on generalizability coefficients, four raters are needed to consistently count relative clauses and three raters to consistently count noun phrases; the same conclusions may be made from the dependability coefficients. It appears not even six raters were enough to count passive voice, complex verbs, and subordinate clauses consistently. Coefficient estimates for construct-irrelevant counts of complex verbs are similar to those for total counts of relative clauses for the mathematics assessments, but much less than for those of other features, suggesting raters were not consistently identifying construct-irrelevant vocabulary. While some grammatical features can be coded consistently by raters, identifying whether these features contain construct-irrelevant vocabulary is more difficult.

Table 2.9.

Decision Study Variance Components and Generalizability Coefficients for the Rater Counts of Grammatical Features: Construct-Irrelevant Grammatical Features on MCAS Mathematics Assessments

Feature	Component	Number of raters				
		2	3	4	5	6
PV	σ_i^2	.070	.070	.070	.070	.070
	σ_r^2	.019	.013	.010	.008	.006
	σ_{ir}^2	.171	.114	.085	.068	.057
	ρ_f^2	.290	.380	.450	.505	.551
	ϕ_f	.269	.355	.423	.479	.524
CV	σ_i^2	.019	.019	.019	.019	.019
	σ_r^2	.004	.003	.002	.002	.001
	σ_{ir}^2	.065	.044	.033	.026	.022
	ρ_f^2	.224	.302	.366	.419	.464
	ϕ_f	.214	.290	.353	.405	.450
SC	σ_i^2	.003	.003	.003	.003	.003
	σ_r^2	.011	.007	.006	.004	.004
	σ_{ir}^2	.227	.151	.113	.091	.076
	ρ_f^2	.014	.021	.028	.034	.041
	ϕ_f	.013	.020	.026	.033	.039
RC	σ_i^2	.381	.381	.381	.381	.381
	σ_r^2	.025	.016	.012	.010	.008
	σ_{ir}^2	.161	.108	.081	.065	.054
	ρ_f^2	.703	.780	.825	.855	.876
	ϕ_f	.672	.754	.804	.837	.860
NP	σ_i^2	2.455	2.455	2.455	2.455	2.455
	σ_r^2	.076	.051	.038	.030	.025
	σ_{ir}^2	.874	.583	.437	.350	.291
	ρ_f^2	.737	.808	.849	.875	.894
	ϕ_f	.721	.795	.838	.866	.886

Note. PV = passive voice, CV = complex verb, SC = subordinate clause, RC = relative clause, NP = complex noun phrase.

Coding Lexical Features

After uncommon words were identified using the VocabProfiler tool, two raters (the author and an undergraduate research assistant) categorized the words as technical or general academic vocabulary independent, using the construct-relevant word lists provided to the raters in the D study. Simple rater agreement was calculated for categorizations for the first assessment rated, the 2018 MCAS Biology assessment, with 85.5% agreement. After resolving discrepancies through discussion, rating continued for the last three assessments, with more favorable agreement: 93.8% agreement on the 2019 MCAS biology assessment, 93.5% agreement on the 2018 MCAS mathematics assessment, and 93.4% agreement on the 2019 MCAS mathematics assessment.

Descriptive statistics across items for each assessment for the total word count, unique technical vocabulary count, unique general academic vocabulary count, and count for words with seven or more letters are in Table 2.10. Assessments on the same subject in different years appear to have similar distributions of these lexical features.

Table 2.10.

Descriptive Statistics for Lexical Features on MCAS Assessments

Test	Descriptive Statistic	Total Words	Technical Vocabulary	General Academic Vocabulary	Words ≥ 7 Letters
MCAS Biology 2018	Mean (SD)	72.5 (28.9)	5.9 (3.1)	4.3 (3.5)	20.4 (9.5)
	Min	20	0	0	4
	Max	139	12	17	48
MCAS Biology 2019	Mean (SD)	69.33 (35.9)	4.7 (2.8)	4.2 (2.8)	18.2 (9.3)
	Min	17	1	0	3
	Max	132	13	11	39
MCAS Math 2018	Mean (SD)	55.9 (50.1)	2.2 (1.9)	1.1 (1.2)	9.0 (7.5)
	Min	10	0	0	1
	Max	277	7	4	36
MCAS Math 2019	Mean (SD)	70.5 (56.9)	2.6 (1.9)	1.1 (1.4)	13.2 (13.8)
	Min	12	0	0	2
	Max	257	8	7	61

Linguistic Complexity Factor Analysis

As the assessments measure differing constructs and raters subsequently had different criteria for identifying construct-irrelevant vocabulary in grammatical features, separate sets of CFAs were conducted on the biology and mathematics assessments with the raters' set of construct-irrelevant counts for grammatical features. As described previously, the raters' set of construct-irrelevant counts was calculated by subtracting the construct-relevant counts from the total counts, leaving behind only construct-irrelevant counts of grammatical features. When looking at the bias influencing EBs, only construct-irrelevant vocabulary should be considered as construct-relevant vocabulary is a construct intended to be measured by the instrument (Avenia-Tapper & Llosa, 2015). In order to claim LC is a potential source of bias leading to DIF against EBs, and to argue the LC in items is responsible for systematic bias against EBs, the LC accounted for must be construct-irrelevant. Results for the mathematics set of CFAs are presented first, followed by the biology set of CFAs.

Mathematics Assessments Factor Analysis

First, the fit indices of unidimensional and multidimensional models of LC were compared. Due to problems with the consistency of raters' counts of some grammatical features, multiple multidimensional models were evaluated: a six dimensional model (with factors for passive voice, complex verbs, subordinate clauses, relative clauses, noun phrases, and lexical complexity), a four dimensional model omitting most features with low consistency (with factors for passive voice, relative clauses, noun phrases, and lexical complexity), and a three dimensional model omitting all features with low consistency (with factors for relative clauses, noun phrases, and lexical complexity). To determine if a unidimensional model fits better than a multidimensional model, multiple unidimensional models were created that only included the features in the multidimensional models. In the unidimensional models created, all lexical and grammatical feature counts were modeled as loading onto a single LC factor. Fit statistics for all tested multidimensional and unidimensional models are presented in Table 2.11. Regardless of how many dimensions were selected, the multidimensional model always fit better than the unidimensional model. The three dimensional model was selected as the best-fitting model as it had the best-fitting fit statistics, and the passive voice factor in the four dimensional model showed non-significant rater count indicators and non-significant variance explained (R^2) by passive voice rater count indicators.

Table 2.11.

Fit Statistics for Determining Multidimensionality of Linguistic Complexity for Mathematics Assessments.

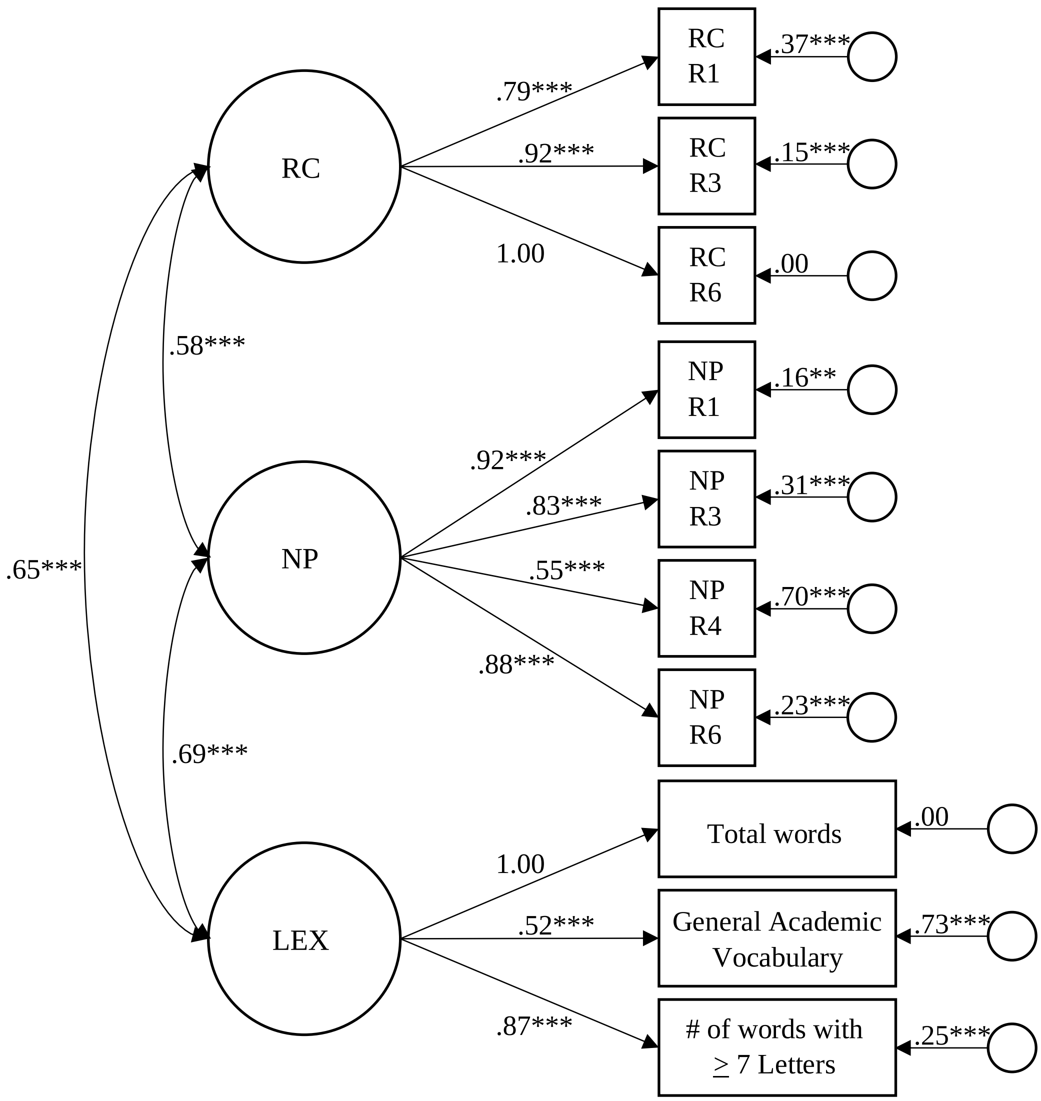
Model	χ^2 (df)	RMSEA	CFI	SRMR
Six dimensions (PV, CV, SC, RC, NP, LEX)				
Multidimensional	411.318*** (194)	.115 [.100, .131]	.789	.112
Unidimensional	679.979*** (209)	.164 [.150, .178]	.543	.109
Four dimensions (PV, RC, NP, LEX)				
Multidimensional	178.265*** (71)	.134 [.110, .159]	.871	.105
Unidimensional	416.835*** (77)	.229 [.208, .251]	.593	.111
Three dimensions (RC, NP, LEX)				
Multidimensional	92.217*** (32)	.150 [.114, .186]	.918	.084
Unidimensional	290.197*** (35)	.295 [.264, .326]	.652	.099

After determining three dimensions should be included in a CFA for raters' counts of LC features, measurement models were created for each factor. The relative clause count model did not include Rater 4 because there was no variance in their construct-irrelevant counts (this rater did not identify any construct-irrelevant relative clauses); this led to the model being just-identified, or having zero degrees of freedom. To determine fit this model, parts of the model had to be constrained. Preliminary models suggested Rater 6's relative clause counts should be fixed to zero and variance fixed to one. The lexical complexity model was also just-identified since there were only three indicators on the factor. Preliminary models suggested fixing "total words" in an item to zero and variance to one. All measurement models fit acceptably well based on Schreiber et al.'s (2006) criteria.

Due to only having three factors, the fit of a model with a higher-order LC factor cannot be tested. Similarly, due to only having two factors for grammatical features, the fit of a model with a higher-order grammatical complexity factor cannot be tested. Therefore, the multidimensional model, with measurement model variations, is the appropriate and best-fitting model for counts of linguistic features in these mathematics assessments. The fit indices are as follows: χ^2 = 98.100 (34), RMSEA = .150 [.116, .185], CFI = .912, SRMR = .099. Factor loadings and correlations are presented in Figure 2.3. In this model, standardized results are presented; all correlations between factors are significant. According to Schriber et al. (2006), model fit is acceptable when RMSEA $\leq$ .08, CFI $\geq$ .95, and SRMR $\leq$ .08. The three dimensional model did not have acceptable fit criteria, and although CFI and SRMR were near acceptable values, RMSEA was well past the threshold of .08.

Figure 2.3.

Multidimensional Model of Linguistic Complexity for the MCAS Mathematics Assessments.



Note. RC = relative clause, NP = complex noun phrase, LEX = lexical complexity, R1 = Rater 1, R3 = Rater 3, R4 = Rater 4, R6 = Rater 6.

Biology Assessments Factor Analysis

Multidimensionality of LC in the biology assessments was evaluated the same way as the mathematics set of models. Results are presented in Table 2.12. Regardless of how many dimensions were selected, the multidimensional model always fit better than the unidimensional model. The three dimensional model was selected as the best-fitting model as it had the best-fitting fit statistics, and the passive voice factor in the four dimensional model showed non-significant rater count indicators and non-significant variance explained (R^2) by passive voice rater count indicators. However, due to three factors leading to a just-identified CFA leaving me unable to explore if a higher-order factor was better fitting than a multidimensional model, higher-order models with four factors were explored.

Table 2.12.

Fit Statistics for Determining Multidimensionality of Linguistic Complexity for Biology Assessments.

Model	**χ^2 (df)**	**RMSEA [90% CI]**	**CFI**	**SRMR**
Six dimensions (PV, CV, SC, RC, NP, LEX)				
Multidimensional	407.401*** (215)	.100 [.085, .114]	.779	.101
Unidimensional	760.197*** (230)	.160 [.148, .173]	.391	.135
Four dimensions (PV, RC, NP, LEX)				
Multidimensional	162.355*** (84)	.102 [.078, .125]	.854	.086
Unidimensional	361.458*** (90)	.183 [.164, .203]	.496	.130
Three dimensions (RC, NP, LEX)				
Multidimensional	86.595*** (41)	.111 [.078, .144]	.890	.081
Unidimensional	197.854*** (44)	.197 [.170, .225]	.629	.099

After determining three or four dimensions should be included in a CFA for raters' counts of LC features, measurement models were created for each factor. The RMSEA fit statistics for the passive voice count and relative clause count models suggested there are issues with fit; the relative clause count model's CFI was below Schreiber et al.'s (2006) recommendation. One of the raters' counts for the relative clause count model had a non-significant loading; this likely contributed to the poor fit indices for this model as this rater under-counted construct-irrelevant relative clauses based on mean values (Table 2.1; $\bar{X}_{R1}$= .411, $\bar{X}_{R3}$ = .622, $\bar{X}_{R4}$ = .089, and $\bar{X}_{R6}$ = .433). The lexical complexity model was just-identified since there were only three indicators on the factor. Preliminary models suggested fixing "Number of words with ≥ 7 letters" in an item to zero and variance to one. The noun phrase and lexical complexity measurement models fit acceptably well based on Schreiber et al.'s (2006) criteria.

Because having three factors means the fit of a model with a higher-order LC factor cannot be tested, the presence of a higher-order LC factor was tested for the four dimension model (Table 2.12). As the higher-order LC model was not better fitting than the multidimensional model, the multidimensional model was retained. Due to there being only three factors for the count of grammatical features, the fit of a model with a higher-order grammatical complexity model could not be tested.

Table 2.13.

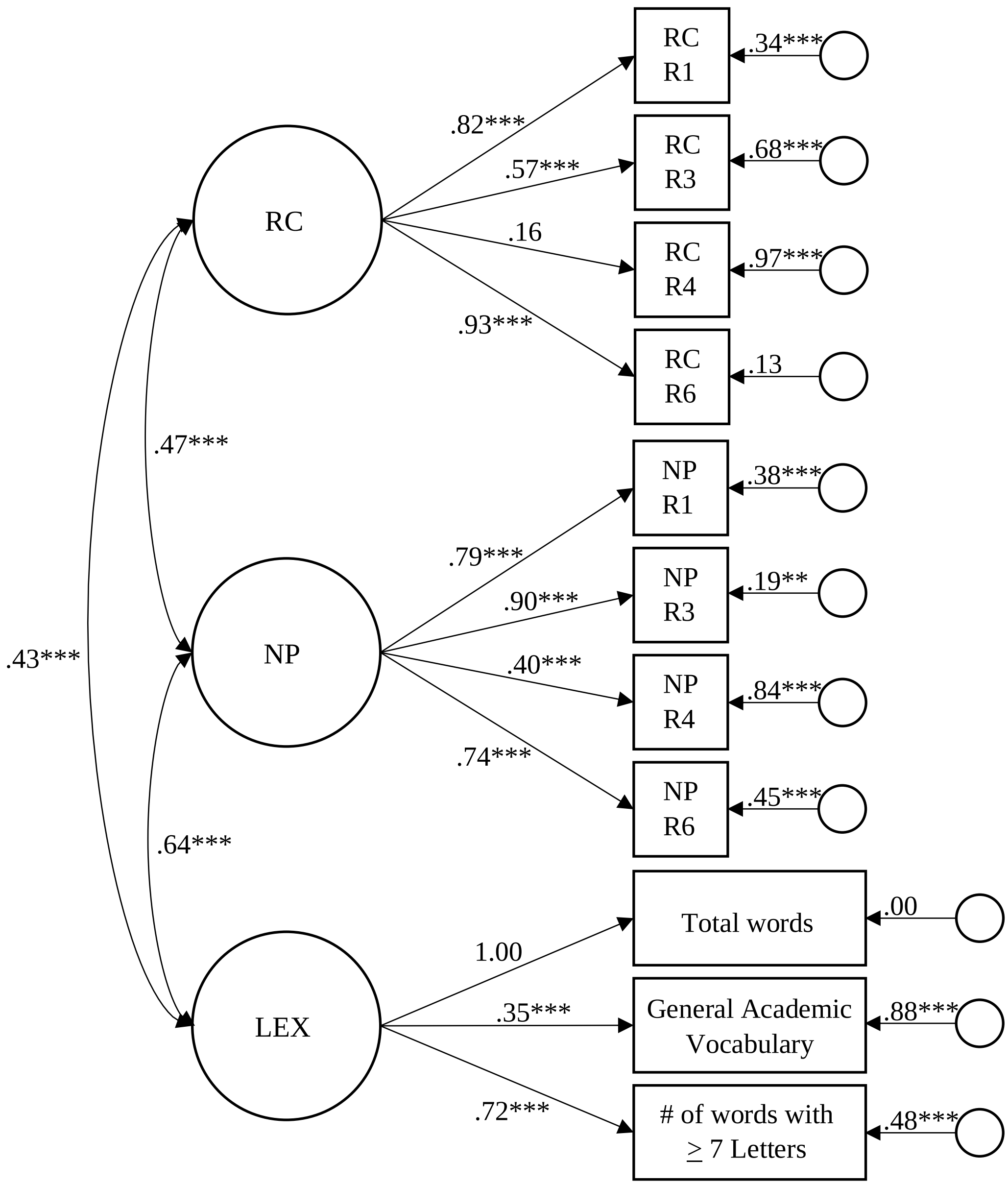
Fit Statistics for Determining Higher-Order Linguistic Complexity Factor for Biology Assessments.

Model	**χ^2 (df)**	**RMSEA**	**CFI**	**SRMR**
Four dimensions (PV, RC, NP, LEX)				
Multidimensional, with measurement model variations	163.461*** (85)	.101 [.078, .125]	.854	.089
Higher-order LC	167.925*** (87)	.102 [.078, .125]	.850	.094

As the three dimensional model had better fit than the four dimensional model (Table 2.13), the three dimensional model, with measurement model variations, is the appropriate and best-fitting model for counts of linguistic features in these mathematics assessments. The fit indices are as follows: χ^2 = 87.579 (42), RMSEA = .110 [.077, .142], CFI = .890, SRMR = .087. Factor loadings and correlations are presented in Figure 2.4. In this model, standardized results are presented; all correlations between factors are significant. Although “Number of words with $\geq$ 7 letters” factor loading was constrained to one and variance to zero, the standardized results applied this constraint to “Total words” instead. According to Schriber et al. (2006), model fit is acceptable when RMSEA $\leq$ .08, CFI $\geq$ .95, and SRMR $\leq$ .08. The three dimensional model did not have acceptable fit criteria, and although fit criteria were somewhat near acceptable values.

Figure 2.4.

Multidimensional Model of Linguistic Complexity for the MCAS Biology Assessments.



Note. RC = relative clause, NP = complex noun phrase, LEX = lexical complexity, R1 = Rater 1, R3 = Rater 3, R4 = Rater 4, R6 = Rater 6.

Discussion

Which Grammatical Features Can Be Consistently Counted?

In the present study, raters had difficulty consistently counting grammatical features in high school biology and mathematics assessments, some raters were under-identifying features. This under-identification is likely due to the raters not being linguistic complex experts, which suggests more rigorous training may be needed if intended to use non-content experts to identify grammatical features, or linguistic content experts would be better raters on this task. For the biology assessments, raters were fairly consistent in their counts of grammatical features, although this consistency decreased when raters had to determine which features included construct-relevant vocabulary. For the mathematics assessments, raters were fairly consistent counting passive voice, relative clauses, and complex noun phrases, but not complex verbs and subordinate clauses. When raters had to determine which features included construct-relevant vocabulary, raters could no longer consistently count passive voice instances, although counting relative clauses and noun phrases was still consistent. Examinations of the variance and covariance components for both subjects revealed that for passive voice, complex verbs, subordinate clauses, and relative clauses, the largest sources of variation were due to raters (ranging from 46.1% to 50.6% for total counts and construct-relevant counts) and the item by raters interaction, or error (ranging from 45.0% to 49.0% for total counts and construct-irrelevant counts). With such a large amount of variation attributable to raters, the training provided to raters must have not sufficiently taught these non-content experts how to identify grammatical features. Alternatively, raters without a background in linguistics may not be able to identify these features systematically without extensive training. The large amount of variation attributable to the item by rater interactions (or error, the “leftover” variance), suggests there is some element of counting grammatical features in items that was not captured. However, when counting noun phrases for either the biology or mathematics assessments, items were the largest source of variation for total counts (38.8% for biology assessments and 48.1% for mathematics assessments), and were still substantially large sources of variation for construct-relevant counts (25.5% for biology assessments and 25.6% for mathematics assessments). Noun phrases were likely easier to identify for raters than these other features, although raters and the item by rater interactions were still large sources of variation in counts.

In most cases, the variation attributed to items decreased between the total count of features and construct-irrelevant count of features. This decrease can be explained by raters relying on (and instructed to use) their total counts to report their construct-relevant counts. For example, Rater 3 identified three instances of passive voice in an item, and Rater 4 identified two. Rater 3 identified one instance (of three) of the passive voice containing construct-relevant vocabulary, whereas Rater 4 identified two instances (out of two). As the initial counting of passive voice was different between raters, this will always influence the count of passive voice containing construct-relevant vocabulary as raters are asked to make secondary judgements based on their primary judgement about the presence of grammatical features in items. This finding was also found for noun phrases, although much more variation in the counts of features was attributed to items than for other features.

Raters Required to Consistently Count Grammatical Features

The results of the decision study demonstrate the need for more raters to achieve a consensus on the counts of construct-irrelevant grammatical features in either subject. On the biology assessments, six raters would be needed to count complex verbs and five raters to count noun phrases consistently, although coefficients for passive voice, subordinate clauses, and relative clauses were still above .700 for six raters, indicating somewhat consistent counting. On the mathematics assessments, four raters would be needed to count relative clauses and three raters to count noun phrases consistently, however coefficients for passive voice, complex verbs, and subordinate clauses were below .600 for six raters, indicating raters had poor consistency counting these features. Due to how low the generalizability coefficients were for the mathematics assessments, the mathematics set of CFAs omitted complex verbs and subordinate clauses from the initial specification, but the factor for passive voice needed to be dropped due to poor fit. The final multidimensional CFA for the biology assessments only included factors for relative clauses, noun phrases, and lexical complexity; to improve fit, the factors for passive voice, complex verbs, and subordinate clauses were removed.

Concerning the model fit of the conducted factor analyses, RMSEA problems may be due to low sample size and low degrees of freedom. Some researchers (Kline, 2023; Jackson, 2003) recommend following the *N:q* rule, or the ratio between the number of cases (*N*) to the number of estimated parameters (*q*). A larger *N:q* ratio is desirable for reliable results, and researchers should aim for a ratio of 20:1. Larger ratios are more likely to have issues with model fit, although many CFAs and structural equation models in published papers have larger ratios than 20:1 (Kline, 2023). The *N:q* ratio was largest in the three dimensional models; there were 23 parameters estimated for 84 mathematics items and 25 parameters estimated for 90 biology items, falling well below the 20:1 ratio recommended. If more items were rated by including items from more assessments from other years of the MCAS mathematics and biology assessments, smaller *N:q* ratios could be obtained, which may improve model fit.

Although there were a large number of degrees of freedom in the starting models with six dimensions (for the multidimensional model, there were 194 degrees of freedom in the mathematics assessment and 215 degrees of freedom in the biology assessment), as variables and factors were removed due to poor model fit, the degrees of freedom became smaller. The final models selected, the three dimensional models, had 32 degrees of freedom in the mathematics assessment and 41 degrees of freedom in the biology assessment. Kenny et al. (2015) investigated the effect of having small degrees of freedom on RMSEA by conducting Monte Carlo simulations of correctly specified models varying by sample size and degrees of freedom. The authors found the percentage of rejected models (models with RMSEA above a cut-off of .10) increased as degrees of freedom or sample size decreased. However, decreasing the degrees of freedom or sample size also increased bias introduced to the sample mean, influencing RMSEA. Taasoobshirazi & Wang (2016) extended this work by confirming this finding for RMSEA, but also explored the effects of small degrees of freedom and sample size on SRMR, CFI, and TLI. The authors found SRMR, CFI, and TLI were not influenced by small sample size or degrees of freedom. The model fit for the mathematics and biology assessments multidimensional models is concerning as the models did not have acceptable model fit based on Schreiber et al.'s (2006) cutoff criteria; this may influence whether the factor scores for relative clauses, complex noun phrases, and lexical complexity (extracted from the measurement models) are predictors of group differences in item responses in the study presented next chapter, although raters with low factor loadings who were less consistent in their counts in the generalizability study will contribute less to the factor score than raters with higher factor loadings.

Abedi et al. (2010) encountered similar problems with some raters that did not accurately identify grammatical features. In their study developing a rubric for measuring the accessibility of reading assessments for students with disabilities, seven raters were used to count the grammatical features in 490 items. Raters were from the applied linguistic department at a university or had backgrounds teaching English to EBs; more specific characteristics about the raters was unavailable. Despite being content experts, the authors found some raters needed more training as raters had different levels of familiarity with each grammatical feature. Out of seven raters, three raters were ready to rate after training, two raters needed training to clarify rating guidelines (for complex noun phrases, although in the report for this study, the authors acknowledged they adjusted the rating guidelines for this feature, leading to confusion for some raters), and two raters struggled with understanding specific features (passive voice and complex verbs). Of these last two raters, one rater was dropped altogether because additional training did not produce results consistent with other raters; this rater continued to undercount features despite repeated training. Each item's grammatical features were counted by two of the six raters, randomly assigned. The results of the present study demonstrates a need for content experts in measuring LC. Counting grammatical features is not a skill that can be quickly taught to others. I suspect specific grammatical features were under-identified which is likely why passive voice, complex verbs, and subordinate clauses were not found fit well in the multidimensional models examined; these same features had less reliability compared to other grammatical features.

Modeling Linguistic Complexity in Assessment Items

It should be noted the final models selected are only generalizable to these MCAS assessments as only MCAS assessments were included in the rating process, although other content assessments using similar design principles may be expected to have similar results. Other features not measured in the present study may contribute to how linguistically complex an item may appear to test-takers. Solano-Flores et al. (2013) described how there are multiple ways to convey meaning in assessment; students do not only rely on the words in test items for meaning-making but incorporate many semiotic features together to make meaning. The features identified by the authors were organized into five modalities: notation (the signs used in mathematics such as abbreviations for units of measurement and symbols for mathematical operation), mathematics register (vocabulary specific to mathematical concepts such as types of numbers and parts of a fraction), natural/mathematical language (mathematical vocabulary used in everyday language such as spelled out units of measurement and numbers), testing register (language used in mathematical assessment and not everyday classroom discourse such as question phrases and comparative phrases), and visual representation (representations without text such as geometric shapes and number lines). Mathematics and natural/mathematic language may contribute to the lexical complexity of an item, although this vocabulary would be construct-relevant and would not be a source of bias against EBs as presumably this vocabulary is what test-takers are expected to know when taking content assessments. However, testing register may contribute to grammatical complexity as the features in this modality relate to the phrases that appear in items such as "which of the following," "equivalent decimal number," and "how many more […] than […]?" (p. 151). While notation and visual representations may appear to mitigate the effects of LC for EBs, Solano-Flores et al. (2013) highlight notation may vary across cultures and limited research exists on how EBs interpret visual representation, although some research suggests the diverse characteristics of EBs may influence their interpretation of visuals on an assessment.

Researchers tend to look at the effect of LC on item responses holistically; when LC is manipulated in a study, items are made more or less linguistically complex by manipulating both grammatical and lexical features (Plath & Leiss, 2018; Riccardi et al., 2020). Treating LC as a unidimensional construct makes it unclear to determine whether test-takers are influenced more by grammatical or lexical complexity. In the present study, evidence was found for both subjects that LC is not a unidimensional construct. This supports the results found by Tomblin and Zhang (2006), who found linguistic complexity to be multidimensional as the age of students increased. Tomblin and Zhang examined whether the dimensionality of language changes as student age by comparing CFA models modeling students' latent language ability. The authors compared a one-factor language model to a two-factor vocabulary-grammar model (this is similar conceptually to the present study's two-factor grammatical and lexical complexity model). They found that the one-factor models tended to fit better for kindergarteners, second graders, and fourth graders, but the two-factor model fit better for the eighth graders. In addition, the correlations between the vocabulary and grammar factors tended to decrease the older students were ("for kindergarten, $r = .941$; for second grade, $r = .934$; for fourth grade, $r = .902$; and for eighth grade, $r = .782$," p. 1201), suggesting latent language ability may be multidimensional. As students use their language ability to decode items of varying linguistic complexity, if language ability is multidimensional, then the linguistic complexity in test items a student must interpret may be multidimensional too. Further research might evaluate the link between multidimensional language ability of test-takers and multidimensional linguistic complexity of items.

Although no evidence for a higher-order grammatical complexity factor was found, the factor scores from relative clause and complex noun phrase counts can be used as proxies for grammatical complexity to determine whether these grammatical features influence item responses. Factor scores from the lexical complexity model were also extracted for use in the next study. Study 2 will use these factor scores to determine how accounting for grammatical features and lexical complexity influences differences in item responses between EBs and

English proficient students. Some prior research suggests lexical complexity may influence item responses more than grammatical complexity, although prior research was conducted with students that were not in high school. Barrot (2013) examined the lexical and syntactic (grammatical) features of texts given to students in second, fourth, and sixth grade to assess their reading comprehension. Barrot found the texts intended for higher grade levels had more lexical features, but there appeared to be an "erratic pattern" (p. 13) for syntactic features unrelated to grade level. The author concluded lexical complexity affects reading comprehension and that syntactic features may not influence reading comprehension as much.

Relationship Between General Academic Vocabulary and Construct-Relevant Terms

While it is important to determine whether grammatical features and lexical complexity differentially affect EBs' item responses, these differences do not necessarily constitute bias. When measuring a construct, such as biology proficiency, we expect that construct to be measured the same for all groups of test-takers (AERA et al., 2014). When that construct is measured differently between groups, bias is exhibited. Bias, or systematic group differences in item responses, between EBs and non-EBs is commonly explained by differences in English proficiency. For example, on a biology assessment, if non-EBs are measured on science content knowledge and EBs are measured on science content knowledge and English proficiency due to unnecessary linguistic complexity in assessment items, then that assessment is biased against EBs because the measured construct is different for the two groups of test-takers. When a test-taker is influenced by something unrelated to the construct of interest, such as unnecessary linguistic complexity, construct-irrelevant variance is introduced (Young, 2008; Haladyna & Downing, 2004; Abedi, 2002; Messick, 1989).

However, we do want to be cautious about what is and is not construct-irrelevant language in assessments (Avenia-Tapper & Llosa, 2015). This is dependent on what construct we want measured by the assessment. Are we interested in learning if students have science or mathematics content knowledge, or are we also interested in learning if students can interpret complex academic language? If we are interested in the former, then for there to be no bias unfairly influencing EBs more than non-EBs, all language on an assessment minus should influence test-takers similarly, with no differences between EBs and non-EBs on item responses attributable to LC. This means that if LC is a significant predictor of group differences in item responses (when conducting DIF analyses), then all LC is construct-irrelevant variance.

However, if we are interested in whether students can interpret complex academic language in addition to measuring content knowledge, this needs to be a stated purpose for the assessment rather than an assumed one. Realistically, we are interested in measuring students' knowledge of both content and complex academic language, or LC. Avenia-Tapper and Llosa (2015) argue that significant correlations between DIF against EBs and LC only reveal items with high LC are more difficult for EBs, and are not sufficient proof of bias against EBs. To determine whether LC in items leads to systematic differences between EBs and EPs, items must be systematically evaluated for construct-relevance so when correlations between DIF against EBs and LC are evaluated, the construct-irrelevant LC in items is a valid explanation for DIF against EBs. This was accomplished in the present study by removing instances of grammatical features containing construct-relevant vocabulary and not including technical vocabulary as a factor loading for lexical complexity. However, raters had difficulty identifying whether features contained construct-relevant language, as rater consistency in counts of grammatical features decreased when raters had to identify construct-relevant language in features. This lower consistency in accounts likely influenced the model fit of the conducted CFAs, which means the factor scores extracted for construct-irrelevant counts of grammatical features in items may be less precise. This may influence whether grammatical features and lexical complexity are found to be significant predictors of item responses or group differences in item responses in the following study.

Regardless, construct-relevant linguistic features need to be removed from consideration when examining whether LC is a significant source of bias of group differences in item responses when conducting DIF analyses, as these features are of interest to the construct being measured and would not be indicators of bias between test-takers of varying English proficiency. Designers of curriculum and assessment need to consider alignment between these content and complex academic language, although recent efforts in improving testing fairness have led to item writers and assessment and curriculum designers to be trained and aware of the effects of complex academic language on students from historically minoritized groups. If students are not explicitly taught complex academic language as it relates to content knowledge, any assessment with complex academic language not covered in curriculum is biased to favor those proficient in English.

CHAPTER THREE

The Effects of Linguistic Complexity on Item Bias Against Emergent Bilinguals: An Explanatory IRT Approach

Linguistic Complexity as a Source of Differential Item Functioning for Emergent Bilinguals

Linguistic complexity (LC) unrelated to the targeted construct has been identified as a common source of construct irrelevant variance in items flagged for differential item functioning (DIF) between emergent bilinguals (EBs) and English proficient students (EPs, or students not identified as EBs). LC may influence DIF because EBs have more difficulties with reading comprehension, particularly with the academic language present on large-scale assessments. Although most research examining the effect of LC on DIF between EBs and EPs has focused on individual linguistic features (Banks et al., 2016; Haag et al., 2013; Heppt, et al., 2015; Kachchaf et al., 2016; Shaftel et al., 2006; Turkan & Liu, 2012), others have proposed LC should be partitioned into lexical and grammatical complexity (Avenia-Tapper & Llosa, 2015; Lee & Randall, 2011; Wolf & Leon, 2009). Few approaches to evaluating items for linguistically complex features affecting EBs have been psychometrically evaluated.

LC influences the difficulty of items and tasks given to students. As LC increases, so can item and task difficulty. Plath and Leiss (2018) conducted a study where they varied the LC in five mathematical tasks by three difficulty levels; higher difficulty levels included less frequently used vocabulary and more complex grammar structure. The authors found students with low German proficiency correctly solved the tasks at lower or similar frequencies when LC was increased, but this pattern was not found for students with high German proficiency, who solved the tasks at lower or higher frequencies when LC was increased. The authors theorized one of the tasks may have been more difficult when linguistically simplified because of missing information or the task was interpreted superficially by students with high German proficiency.

Although the relationship between LC and DIF between EBs and EPs has been studied, researchers have yet to examine whether accounting for LC in the IRT models used to identify DIF decreases DIF detection and the magnitude of DIF. This could be accomplished through the use of explanatory item response models (EIRM). With EIRM, item-level covariates (such as lexical complexity) can be included into IRT models using a nonlinear mixed model framework to predict the effect of covariates on item difficulty (De Boeck & Wilson, 2004).

Past research has looked at the relationship between LC features and item difficulty, but without directly examining the differential effect LC may have on item responses for EBs and EPs. Wolf and Leon (2009) examined the relationship between linguistic rating scores and DIF between EPs and different subgroups of EBs (all EBs, high English proficiency EBs, and low English proficiency EBs). They conducted correlation analyses between linguistic rating scores and DIF statistics for "easy" (75% of EPs answered the item correctly) and "not-easy" items. For easy items across all comparison groups, the authors found significant correlations between DIF favoring EPs and total words, academic vocabulary (general academic and technical), form ("proportion of language to nonlanguage in an item," p. 144), and reliance (language knowledge needed to correctly answer an item). However for the "not-easy" items there were less consistent patterns; across all comparison groups, the authors found significant correlations between DIF favoring EPs and reliance and DIF favoring EBs and technical vocabulary. For the low English proficiency EB group, no other correlations between linguistic rating scores and DIF statistics were significant; for the high English proficiency EB group, number of sentences and cohesion (linguistic devices that "connect text within or across clauses," p. 144) were significant.

In addition, Wolf and Leon found science tests tended to have higher linguistic demands than mathematics tests and tests from higher grades had higher linguistic demands than tests from lower grades. They found more DIF items detected in science tests than math tests and more DIF items were detected when the focal group was low English proficiency EBs and not high English proficiency EBs. The authors suggest looking at EB students' opportunity to learn ("uncovering the ways that ELL students are exposed to and instructed on both general and specific academic language," p. 156). EBs have different opportunities to learn compared to their monolingual and reclassified as English proficient peers; teachers have lower expectations for these students than their high-tracked students, both linguistically and academically (Callahan, 2005). In Callahan's (2005) study exploring the effects of ability tracking on the academic outcomes of EBs, teachers reported expecting less academically from their lower English proficiency EBs, many of which were recent immigrants, compared to their higher English proficiency EBs. If EBs have different opportunity to learn based on subgroup characteristics, these will affect their item responses on assessments.

Study Hypotheses

There are three specific hypotheses for this study. Each hypothesis presented is followed by rationale:

1. LC factor scores will have significant main effects and interactions with emergent bilingual status; the interactions will favor English proficient students.
2. For items with higher LC, there will be less items flagged as significantly favoring EPs when including LC as a covariate.
3. For items with lower LC, there will be no change in items flagged as significantly favoring EPs when including LC as a covariate.

Hypothesis 1 is drawn from Wolf and Leon (2009). They found items that were easier tended to exhibit larger magnitudes of DIF and the authors speculated that this had to do with the higher amounts of LC in the items. LC was not accounted for in the models used in this study. If LC is the main contributor of DIF between EBs and EPs, the interaction between LC factor scores and emergent bilingual status should favor English proficient students. It is expected that when including LC as a covariate in DIF analyses (discussed in the Methods section), the main effect of LC on item responses will be significant. If LC has significant main effects on item responses for test-takers, test developers need to consider if the language in items is a construct they want to measure, and consider using a range of LC with lower LC and higher LC items on their assessments. However, if interactions of LC features with EB status are significant, scores between groups should be interpreted with caution, as the LC in items is influencing test-takers in these groups differently, which may introduce bias.

The present study aims to investigate the interactions of LC covariates and EB group membership to determine if the LC in test items differentially affects the item responses of EBs compared to EPs, which leads to the next two hypotheses. Hypotheses 2 and 3 are also drawn from Wolf and Leon (2009). Items with significant DIF and higher LC are expected to favor EPs per their results. When LC is accounted for, then items with higher LC that favor EPs should either exhibit non-significant DIF or favor EBs, as the assumed source of DIF is accounted for. If items significantly favor EBs and have higher LC, accounting for LC should not change the direction or significance of DIF because LC is not the expected source of DIF in these items. Items with significant DIF favoring EPs and lower LC are not expected to change DIF direction or significance because the source of DIF (some factor that isn't LC) was not accounted for. By accounting for LC in models, less DIF should be captured as the potential bias introduced by LC in items is accounted for. This also allows test developers to explore other sources of bias in items if DIF is present after accounting for LC. These study hypotheses can be evaluated with explanatory item response models (EIRMs), an extension of IRT modeling.

Explanatory Item Response Models

To begin discussing EIRMs, first Rasch models need to be examined. A Rasch model is a one-parameter logistic model used to model binary item responses in IRT. A simplified equation for a Rasch model, Equation 3.1, shows how a person's latent ability, or person parameter, (θ_p) and an item's difficulty, or item parameter, (β_i) can be used to predict η_{pi}, or the natural log of ($p/(1-p)$), where p is the probability a person's response to an item is correct given that person's latent ability (De Boeck & Wilson, 2004; Fischer, 1973; Rasch, 1960). The Rasch model can be extended to include person and item predictors as well as polytomous items, or items that are scored across more categories than correct and incorrect.

$$\eta_{pi} = \theta_p - \beta_i \tag{3.1}$$

The Rasch model in Equation 3.1 can be conceptualized in a multi-level format using hierarchical generalized linear modeling (HGLM) framework. Kamata (2001) demonstrated how the Rasch model can be specified as a two-level hierarchical linear model with item responses nested within persons. Equation 3.2 shows the result of the derivation of the level-1 or item-level models. In this model, η_{ij} still represents the natural log of the probability a person's response to an item is correct given that person's latent ability; however, variables for person j are now included. To use this model with an assessment with k items, one item is set as the reference or anchor item, and k – 1 item coefficients are calculated. β_{0j} is interpreted two ways, as the effect of the reference item for person j, or the latent trait estimate for person j (Pastor, 2003). The dataset for this model is prepared in long form, with a row for a person's response to a particular item. Variables for item indicators are included in this dataset. X_{qij} is the item indicator for person *j* for item *i*. X_{qij} = 1 when *q* = *i* and 0 when *q* ≠ *i*. β_{qj} (β_{1j} through $\beta_{(k-1)j}$) are the effects of the *q*th item compared to the reference item, when *q* = *i*.

$$\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} \tag{3.2}$$

The level-2 or person-level models derived by Kamata (2001) are presented in Equation 3.3. In this equation, random component u_{0j} is added to the intercept to show how the abilities of test-takers vary across persons, but not across items. The random component u_{0j} is normally distributed with a mean of 0 and τ variance. β_{0j} is decomposed into the γ_{00}, the effect of the intercept (difficulty for the reference item) and u_{0j}, the random component representing person *j*'s ability. β_{qj} (β_{1j} through $\beta_{(k-1)j}$) are the effects of the *q*th item compared to the reference item and γ_{q0} (γ_{10} through $\gamma_{(k-1)0}$) are the mean effects of the *q*th item (equivalent to β_{qj} with item effects). With these effects, the item difficulties for each item can be calculated as γ_{q0} + γ_{00}.

$$\begin{gathered} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \vdots \\ \beta_{(k-1)j} = \gamma_{(k-1)0} \end{gathered} \tag{3.3}$$

DIF is detected in an item when members of different groups with the same underlying latent ability have different probabilities of responses to that item (De Ayala, 2022). In DIF analyses, two groups of test-takers are compared: a focal group, typically a marginalized or underrepresented group of test-takers in education, such as EBs, and a reference group, which can be a control group or a group of test-takers with more representation or privilege, such as

EPs. When an item's difficulty is found to be significantly different between the focal and reference groups, uniform DIF is present for that item. If an item is significantly more difficult for the focal group than for the reference group, the item is biased against the focal group. If an item is significantly less difficult for the focal group than for the reference group, the item is biased in favor of the focal group.

DIF between groups of test-takers (such as EBs and EPs) can be evaluated by including person characteristics in the level two models of the Rasch HGLM (Van den Noortgate & De Boeck, 2005). Some advantages of using Rasch HGLMs for assessing DIF include estimating multiple items for DIF simultaneously, including dichotomous and polytomous items, and adding item features as covariates to explain DIF (Chen, et al., 2013). This model is represented in Equation 3.4. Now β_{0j}, the intercept for the reference item and latent trait estimate for person j, is decomposed into the γ_{00}, the effect for the intercept (reference item's difficulty for the reference group), γ_{01}, the main effect of belonging to focal group G_j (the difference in ability between a reference group test-taker and a focal group test-taker), and u_{0j}, the random component representing person j's ability. γ_{00} represents the latent ability estimate for a reference group test-taker and γ_{01} is the effect of focal group status on latent ability controlling for other variables in the model, with u_{0j} as the as the latent trait estimate for person j after controlling for the effect of focal group status (Pastor, 2003). G_j is 1 when the person belongs to the focal group (typically the historically underrepresented group, in this study this would be emergent bilinguals) and 0 when the person belongs to the reference group (in the present study, this would English proficient students). Positive values of γ_{01} signify test-takers in the focal group had higher ability estimates; negative values of γ_{01} signify test-takers in the reference group had higher ability estimates (Ravand, 2015).

β_{qj}, the effects of the *q*th item compared to the reference item can be similarly be decomposed into γ_{q0} (γ_{10} through $\gamma_{(k-1)0}$), the effects of the *q*th item compared to the reference item for the reference group and γ_{q1} (γ_{11} through $\gamma_{(k-1)1}$), the effect of belonging to focal group G_j for the *q*th item. γ_{q0} is the mean item effect, γ_{q1} represents the DIF on item *q* above the DIF introduced by γ_{01}, or the differences in item difficulty associated with focal group status (Williams & Beretvas, 2006). When γ_{q1} is significant, that item exhibits DIF, or group differences in responding to that item. Positive values of γ_{q1} signify the item favors the focal group; negative values of γ_{q1} signify the item favors the reference group. This is in relation to γ_{01}, which shows group differences in responding to te reference item. As γ_{q1} is interpreted as the difference to the reference item, an adjusted DIF estimate needs to be calculated that considers the effects of both γ_{q1} and γ_{01}. Criteria for significant DIF will be discussed in the methods section of this chapter. For the reference group (G_j = 0), item difficulties are calculated as $\gamma_{q0} + \gamma_{01}*0 + \gamma_{00} + \gamma_{11}*0$, which reduces to $\gamma_{q0} + \gamma_{00}$. For the focal group ($G_j$ = 1), item difficulties are calculated as $\gamma_{q0} + \gamma_{01}*1 + \gamma_{00} + \gamma_{11}*1$, which reduces to $\gamma_{q0} + \gamma_{01} + \gamma_{00} + \gamma_{11}$.

$$\eta_{ij} = \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij}$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(G_j) + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(G_j) \\ &\vdots \\ \beta_{(k-1)j} &= \gamma_{(k-1)0} + \gamma_{(k-1)1}(G_j) \end{aligned} \tag{3.4}$$

However, the models introduced thus far are not EIRMs. To examine the effects of LC on item responses, an EIRM needs to be used, specifically an item explanatory model (De Boeck &

Wilson, 2004). This can be accomplished with a linear logistic test model (Equation 3.5), an extension of the Rasch model that expands β_i in Equation 3.1 to consider the effects of multiple, or *k* item properties, on η_{ij} instead of one item property (item difficulty). The effects of the item properties (β_k) are influenced by the value of the item property (X_{ik}).

$$\eta_{ij} = \theta_j - \sum_{k=0}^{K} \beta_k X_{ik} \tag{3.5}$$

Kamata's (2001) work has been extended since the original reformulation of the Rasch model as an HGLM. Additional regression coefficients based on explanatory item covariates can be added to the model as a level-2 predictor (De Ayala, 2022; Pettersen & Braeken, 2019; Janssen, et al., 2004; Swanson et al., 2002). Consider in the base HGLM, each item indicator is an item characteristic, the property of belonging to item *i*; other item covariates can similarly be included. Thus, the model in Equation 3.5 can be modified to incorporate item characteristic *s* in order to examine the main effect of item characteristic *s* and the interaction between item characteristic *s* and focal group belonging (Equation 3.6). In this model, Y_{sqi} is the value of item characteristic *s* for the *q*th item; in the present study this would be the inclusion of a linguistic complexity factor score. Y_{sqi} is the value of the item characteristic *s* for the *q*th item when $q = i$ and 0 when $q \neq i$. β_{sj} is the effect of the *s*th item characteristic *s* for person *j*, this can be decomposed into γ_{s0}, the main effect of item characteristic *s* on item difficulty, γ_{s1}, the effect of item characteristic *s* and belonging to focal group G_j, and u_{sj}, the random effect of item characteristic *s* on person *j*.

As in Equation 3.4, β_{0j}, the intercept for the reference item, is decomposed into the γ_{00}, the effect for the intercept (reference item difficulty), γ_{01}, the main effect of belonging to focal group G_j (the difference in ability between a reference group test-taker and a focal group test-taker) after controlling for all other variables, and u_{0j}, the random component representing person *j*'s ability. Recall that X_{qij} is the item indicator for person *j* for item *i; X_{qij}* = 1 when *q* = *i* and 0 when $q \neq i$. β_{qj}, the effects of the *q*th item are decomposed into γ_{q0} (γ_{10} through $\gamma_{(k-1)0}$), the effects of the *q*th item compared to the reference item, and γ_{q1} (γ_{11} through $\gamma_{(k-1)1}$), the effect of belonging to focal group G_j for the *q*th item. When γ_{q1} is significant, that item exhibits DIF, or group differences in responding to that item, after conditioning for item characteristic *s*. Positive values of γ_{q1} signify the item favors the focal group; negative values of γ_{q1} signify the item favors the reference group.

$$
\begin{aligned}
\eta_{ij} &= \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(G_j) + u_{0j} \\
\beta_{sj} &= \gamma_{s0} + \gamma_{s1}(G_j) + u_{sj} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}(G_j) \\
&\vdots \\
\beta_{(k-1)j} &= \gamma_{(k-1)0} + \gamma_{(k-1)1}(G_j)
\end{aligned}
\tag{3.6}
$$

Often times, assessments include polytomous items. Williams and Beretvas (2006) demonstrated how Kamata's (2001) Rasch HGLM can be extended to incorporate polytomous items with *m* categories and corresponds to a rating scale model. In a rating scale model, thresholds are constrained to be the same for all items. Rating scale models can incorporate continuous item covariates unlike partial credit models, which can only incorporate item-by-category covariates (Rijmen et al., 2003). This means continuous item covariates, like the factor scores for lexical complexity, complex noun phrases and relative clauses obtained in Study One, can be incorporated into a rating scale model.

For each category, there are m – 1 sets of level-1 equations to obtain estimates for item thresholds. Equation 3.7 demonstrates how a Rasch HGLM that consider DIF and item characteristic *s* with items with up to five categories (the maximum number of categories for an item in the present study) would be formulated, analogous to a rating scale mode. Multiple level one models are considered based on the probability of scoring in a particular category; in the Rasch HGLM, η_{ij} denotes probability of a correct response for dichotomous items. In this polytomous Rasch HGLM, η_{0ij} denotes the probability of scoring one point, η_{1ij} denotes the probability of scoring two points, η_{2ij} denotes the probability of scoring three points, η_{3ij} denotes the probability of scoring four points. For polytomous items, item thresholds are calculated instead of item difficulties; these thresholds represent where there is a 50% chance for test-takers to score in adjacent categories *m* and *m* – 1 (Eckes, 2015). In the level one equations, threshold parameters for the differences between thresholds are also included; δ_1 is the threshold difference between scoring one and two points, δ_2 is the threshold difference between scoring two and three points, and δ_3, is the threshold difference between scoring three and four points. Group differences in these threshold parameters were also evaluated in the present study; thresholds were treated as missing for dichotomous items.

The only differences between η_{0ij}, η_{1ij}, η_{2ij}, and η_{3ij} are the threshold differences. β_{0j}, β_{sj}, and each β_{qj} are the same level two equations for each level one model. As in Equation 3.4, β_{0j}, the intercept for the reference item, is decomposed into the γ_{00}, the effect for the intercept (reference item difficulty) or estimate of ability for a reference group test-taker, γ_{01}, the main effect of belonging to focal group G_j (the difference in ability between a reference group test-taker and a focal group test-taker) on latent ability estimates, and u_{0j}, the random component representing person *j*'s ability. X_{qij}, the item indicator for person *j* for item *i*, is equal to 1 when *q*

= i and 0 when $q \neq i$. β_{qj}, the effects of the qth item are decomposed into γ_{q0} (γ_{10} through $\gamma_{(k-1)0}$), the effects of the qth item compared to the reference item, and γ_{q1} (γ_{11} through $\gamma_{(k-1)1}$), the effect of belonging to focal group G_j for the qth item. As in Equation 3.5, the effect of item characteristic s can be evaluated, where Y_{sqi} is the value of item characteristic s for the qth item when $q = i$ and 0 when $q \neq i$. β_{sj} is the effect of the sth item characteristic s for person j, this can be decomposed into γ_{s0}, the main effect of item characteristic s, γ_{s1}, the effect of item characteristic s and belonging to focal group G_j, and u_{sj}, the random effect of item characteristic s on person j.

$$\eta_{0ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj} X_{qij}$$

$$\eta_{1ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} + \delta_1$$

$$\eta_{2ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} + \delta_2$$

$$\eta_{3ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} + \delta_3$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(G_j) + u_{0j} \tag{3.7}$$

$$\beta_{sj} = \gamma_{s0} + \gamma_{s1}(G_j) + u_{sj}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(G_j)$$

$$\vdots$$

$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(G_j)$$

$$\delta_1$$

$$\delta_2$$

$$\delta_3$$

Evaluating Subgroups of Emergent Bilinguals for Differential Item Functioning Analyses

When looking at the performance of underrepresented groups on assessments, EBs tend to be examined as a whole. EBs are heterogenous populations with several characteristics that may influence their performance on assessments such as length of time classified as an emergent bilingual, language spoken, student with limited/interrupted formal education status, disability status, type of disability, etc. Overall, limited research exists that examines item response differences between EPs and subgroups of EBs and how LC may affect some subpopulations of EBs more than others.

Lane and Leventhal (2015) summarize important considerations in looking at DIF for EBs and students with disabilities, particularly for EB subgroups and specific categories of disabilities. The factors that influence these DIF analyses include “small sample sizes, nonoverlapping proficiency distributions, and lack of measurement precision” (Lane & Leventhal, 2015, p. 188) which can be accounted for by using large-scale test data and evaluation effect sizes. They suggest nonoverlapping proficiency distributions are because EBs and students with disabilities generally score lower on assessments than EPs and students without disabilities and EBs and students with disabilities may have less access to the construct being measured because of their English proficiency and disability, respectively. The authors argue the heterogeneity of EB students and students with disabilities should be addressed by dividing EBs and students with disabilities into subgroups and examining the efficacy of accommodations for these subgroups as well as the psychometric properties of tests and their items for these subgroups. Lane and Leventhal also highlight the few studies that have examined DIF by subgroup for EBs and students with disabilities.

Notably, Kato et al. (2009) conducted DIF analyses examining students with specific disabilities. They found many items exhibited DIF, only some showed substantive DIF. Their finding was to treat students with disabilities as a heterogenous group and argued to conduct DIF analyses for these subgroups. These small sample sizes may affect statistical values, especially as subgroups get smaller and more precise in their categories. The high heterogeneity of student characteristics of the students with disabilities population and the EB populations may lead to greatly reduced DIF detection rates and lower rates of false positives as found in a simulation study using logistic regression and Mantel-Haenszel DIF methods (Oliveri et al., 2014).

DIF is between subgroups of EBs and EPs has likely not been studied extensively due to the large sample sizes needed to conduct DIF analyses. DIF analyses require large sample sizes (thousands of students), as sample sizes that are too small lead to less accurate results (Sireci et al., 2018). Therefore, test-takers end up grouped into the population they best fit in without considering the heterogeneity of the characteristics of that population in order to meet this sample size requirement. For example, DIF analyses are conducted between EBs and EPs because the generally low level of English proficiency EBs have compared to EPs that may lead to differences in item responses between groups. However, there are varying levels of English proficiency among EBs and EBs come from many differing backgrounds that contributes to their English proficiency. Therefore, these characteristics may also influence item responses. If DIF analyses between subgroups of EBs and EPs are not conducted, this threatens the validity of interpretations made about the abilities of EBs from those assessments, as some subgroups' item response differences may be masked by larger EB subgroups (Faulkner-Bond & Sireci, 2015; Lane & Leventhal, 2015).

The present study will explore the relationship between LC and item responses for different subpopulations of EBs in DIF analyses comparing subgroups of EBs to EPs and between subgroups of EBs. Two characteristics of EBs will be examined: status as a LTEB and whether Spanish is the first language of the EB. These characteristics of EBs may contribute to the role LC plays in EBs' item responses.

Length of Time as an Emergent Bilingual

García et al. (2008) provide a comprehensive summary on how EBs are identified and reclassified, although specific requirements vary by state. In many states, students take a home language survey when they enroll in a new school, if a student identifies a language other than English is spoken at home, the EB is then assessed for English language proficiency. English language proficiency assessments are also commonly used for reclassifying EBs, as are state achievement test performance, teacher referral, and parent referral. However, some EBs take longer to attain reclassification than others. In their report on EB reclassification in New York City public schools, Kieffer and Parker (2016) found that while 52% of EBs who have been enrolled since kindergarten attained reclassification within four years (the expected time to reclassification for many states), 25% of these EBs did not attain reclassification after six years and were considered "long-term emergent bilinguals" or LTEBs. Exact definitions for LTEBs vary, but many researchers agree an EB is considered LTEB when they have spent five or six years enrolled without attaining reclassification (Menken et al., 2012; Olsen, 2010; Olsen, 2014). Kieffer and Parker also found that of the EBs in New York City public schools, 37% of students with below average initial English proficiency became LTEBs compared with 19% of students with above average initial English proficiency; the authors speculated students with lower initial English proficiency may need extra support. In addition, of EBs analyzed, 63% of EBs with specific learning disabilities and 46% of EBs with speech or language impairments became LTEBs.

While increasing reclassification rates is not the goal of this discussion, LTEBs are a group that are facing difficulties attaining reclassification, whether that be due to their difficulties with overall English proficiency, academic English proficiency, or assessment performance. These difficulties do not only affect whether they are reclassified, but it also influences their day-to-day learning in classrooms and performance on state achievement tests. We ought to focus on what barriers and challenges LTEBs are facing, but we also need to recognize the strengths these students have and could have with the right support, such as by implementing policies that support EBs' bilingualism and acquiring of their native language and English, and by shifting away from English-only instruction (García et al., 2008).

LTEBs face inconsistent support and instructional services that are preventing these students from attaining sufficient English proficiency for reclassification (Shin, 2020). Some authors have found language support services for EBs in high school centers on the needs of non-LTEBs, e.g. learning English versus developing academic language (Kim & García, 2015; Menken et al., 2012). Without focusing on academic language, the LC in items challenges EBs; some researchers suggest academic English is an obstacle for LTEBs' reclassification (Brooks, 2015; Menken et al., 2012; Olsen, 2010). This difficult in attaining English proficiency may influence their item responses in a way that may be different than that of EBs not identified as LTEBs. Conducting DIF analyses between EPs and LTEBs and between EPs and non-LTEBs can reveal whether including LTEBs and non-LTEBs in the same group leads to masking of DIF effects. If different items are flagged or if the sign of the DIF coefficient changes direction between DIF analyses, then assessment developers may need to conduct separate DIF analyses by LTEB status or length of time as an EB.

First Language

Spanish is the most common language spoken by EBs in the United States; in Fall 2018, 75.2% of all EBs were identified as Spanish-speakers (NCES, 2021). Consequentially, research examining DIF in content assessments by native language tends to examine DIF presence between EBs and EPs or Spanish-speaking EBs and EPs. This may be because of the sample size required to have enough statistical power to conduct analyses examining specific language groups, or to reduce the variance introduced by an EB speaking a language other than Spanish for a sample including all EBs.

While not looking at DIF, Solano-Flores and Li (2009; 2006) have examined the relationship between EBs, language, and assessment performance by focusing on specific subgroups of EBs. They examined how Spanish-speaking, Haitian-Creole speakers, and Chinese language-speaking EBs' item responses differed based on whether EBs were administered the item in their native language or English. The researchers found differences between groups by item language, which suggests there may be differences in how native language influences responses on an assessment written in English. Solano-Flores (2014) surmised from these studies that the largest source of measurement error was the interaction between students, items, and language/dialect (Solano-Flores & Li, 2009; Solano-Flores & Li, 2006). EBs' performance appears to depend on their strengths and weaknesses in their native language and English and the linguistic challenges of items given in their native language and English. As a result, EBs from different linguistic groups may need to be assessed on differing amounts of items to obtain dependable scores on content assessments (Solano-Flores & Li, 2006).

Keeping in mind the challenges in obtaining accurate findings with a small sample size, the present study partitioned EBs into Spanish-speaking and non-Spanish-speaking EBs. However, readers should note that within the non-Spanish-speaking EB sample, some languages are more dominant than others and may mask the effects of students speaking less common non-English languages. Conducting DIF analyses between EPs and Spanish-speaking EBs and between EPs and non-Spanish-speaking EBs can reveal whether including Spanish-speaking and non-Spanish-speaking EBs in the same group leads to misrepresentation of DIF effects. If different items are flagged or if the sign of the DIF coefficient changes direction between DIF analyses, then assessment developers may need to conduct separate DIF analyses by first language. In addition, the effects of LC on item responses for non-Spanish-speaking EBs may be masked by the larger majority of EBs speaking Spanish. DIF analyses examining the effect of LC on item responses would need to include subgroup comparisons.

Methodology

In this section, the participants and materials used in the study are described. This is followed by the procedures for data preparation and analyses steps.

Participants

Data was drawn from two large-scale assessments: the 2019 10th grade mathematics MCAS and the 2019 high school biology MCAS (DESE, 2019a; DESE 2019b). The Massachusetts Department of Elementary and Second Education (DESE) has released all items for these assessments along with deidentified student-level data for item responses. The MCAS is administered to thousands to students each year, including thousands of EBs; this satisfies the demand for larger samples required for IRT models. In the present study, there are 3,969 EBs and 66,423 EPs in the 2019 high school mathematics MCAS (“mathematics assessment”) sample and 1,922 EBs and 15,214 EPs, or English proficient students (EPs), in the 2019 high school biology MCAS (“biology assessment”) sample. EBs were partitioned into subsamples based on length of time as an EB and first language to evaluate the effects of LC on item responses for subgroups of EBs. If an EB was enrolled in Massachusetts schools for six or more years, they were classified as an LTEB, otherwise they were categorized as "short-term emergent bilingual” or STEB. While STEB is not a label applied to EBs in practice, this paper uses STEB as shorthand to referring to those students who are EBs, but not LTEBs. If an EB was reported as having Spanish as a "first language," they were classified as a Spanish-speaking EB, EBs identified as having another language as their first language were categorized as non-Spanish-speaking EBs. In the mathematics assessment sample, there are 1,274 LTEBs and 2,695 STEBs, and 1,591 non-Spanish-speaking EBs and 2,378 Spanish-Speaking EBs. In the biology assessment sample, there are 504 LTEBs and 1,419 STEBs, and 723 non-Spanish-speaking EBs ("OTH" in tables) and 1,199 Spanish-Speaking EBs ("SPA" in tables). Variables were created for comparison groups, as DIF analyses were conducted for combinations of EPs and EBs and EBs and EBs; Table 3.1 lists the comparison groups used in the present study and the abbreviation used to refer to each comparison group in the results section. The first group in the "Comparison Group Abbreviation" column is the reference group (coded as "0"), and the second group is the focal group (coded as "1").

Table 3.1.

Comparison groups for DIF analyses

Comparison Group Category	Groups Compared	Comparison Group Abbreviation
Baseline	EP vs. EB	EPvEB
Length of time as EB	EP vs. STEB	EPvSTEB
	EP vs. LTEB	EPvLTEB
	STEB vs. LTEB	STEBvLTEB
First language	EP vs. Spanish-speaking EB	EPvSPA
	EP vs. Non-Spanish-speaking EB	EPvOTH
	Spanish-speaking EB vs. Non-Spanish-speaking EB	OTHvSPA

For the mathematics assessment sample, demographic characteristics for EBs, EB subsamples, and EPs are presented in Table 3.2. There appear to be more male students represented in the LTEB subsample than in other subsamples. Compared to EPs (which included reclassified EBs), who are predominantly White, the majority of EBs are Hispanic. Racial differences between EB subsamples emerge between Spanish-speaking and non-Spanish-speaking EBs; Spanish-speaking EBs are nearly 100% Hispanic, with non-Spanish-speaking EBs appearing to be mainly Black and White students. EBs, LTEBs, STEBs, and EPs have similar enrollment distributions to the biology assessment sample, but Spanish-speaking EBs appear to be enrolled in Massachusetts schools longer than non-Spanish-speaking EBs, although average years enrolled in the same district is similar between Spanish-speaking and non-Spanish-speaking EBs. EBs are identified as economically disadvantaged at a rate more than double than that of EPs. Students identified as economically disadvantaged participated in state-administered programs such as food stamps, welfare, foster case, or Medicaid; there was no indicator for participation in the free or reduced-price lunch program. When examining EB subgroups, 46.5% of LTEBs have an IEP compared to 7.3% of STEBs, and 24.6% of Spanish-speaking EBs have an IEP compared to 12.9% of non-Spanish-speaking EBs. EBs were identified as homeless at a much greater rate than EPs, but there were differences in homeless rates by EP sample. LTEBs and Spanish-speaking EBs appear to experience more homelessness than STEBs and non-Spanish-speaking EBs. The distribution of performance level on the mathematics assessment by subsample suggests that across the board, EBs are not meeting Massachusetts assessment expectations, although non-Spanish-speaking EBs have higher rates of proficient and advanced performance levels than other EB subsamples, with 16.1% of non-Spanish-speaking EBs scoring proficient or advanced compared to 4.2% of Spanish-speaking EBs, 10.4% of STEBs, and 6.1% of LTEBs.

Table 3.2.

Demographic characteristics for students by subsample – Mathematics Assessment

Characteristic	EP	EB	STEB	LTEB	OTH	SPA
n	66423	3969	2695	1274	1591	2378
Female	49.5%	45.1%	46.9%	41.2%	45.6%	44.8%
Male	50.4%	54.9%	53.0%	58.8%	54.4%	55.2%
Asian	6.8%	7.4%	8.1%	6.0%	18.5%	0.0%
African-American/Black	8.3%	18.7%	18.0%	20.3%	46.3%	0.3%
Hispanic or Latino	15.2%	64.7%	63.5%	67.0%	14.5%	98.2%
Multiracial, non-Hispanic or Latino	3.3%	0.6%	0.3%	1.2%	1.2%	0.2%
American Indian or Alaskan Native	0.3%	0.2%	0.2%	0.0%	0.1%	0.2%
Native Hawaiian or Pacific Islander	0.1%	0.2%	0.1%	0.2%	0.4%	0.0%
White	66.2%	8.3%	9.8%	5.3%	19.0%	1.2%
Avg. years student attended MA schools	10.0	5.0	2.9	9.6	4.5	8.0
Avg. years student continuously enrolled in district	7.2	3.8	2.5	6.6	3.5	4.1
Economically Disadvantaged	27.4%	74.3%	73.8%	75.4%	67.9%	78.6%
IEP Status	17.0%	19.9%	7.3%	46.5%	12.9%	24.6%
Homeless	1.2%	9.6%	11.6%	5.3%	5.2%	12.6%
Advanced	14.0%	0.8%	1.2%	0.0%	1.8%	0.1%
Proficient	47.6%	8.2%	9.2%	6.1%	14.3%	4.1%
Needs improvement	31.6%	48.7%	48.1%	50.0%	51.2%	47.0%
Failing	6.8%	42.3%	41.6%	43.9%	32.6%	48.8%

In Table 3.3, the twelve most common first languages for students in the mathematics assessment sample are presented by subsample. The distributions of first languages between LTEBs and STEBs is similar, although there appear to be higher rates of Spanish-speaking LTEBs than STEBs and Portuguese and Chinese-speaking STEBs than LTEBs.

Table 3.3.

First languages for students by subsample – Mathematics assessment

First language	EP	EB	STEB	LTEB	OTH	SPA
n	66423	3969	2695	1274	1591	2378
English	86.1%	-	-	-	-	-
Spanish	6.1%	59.9%	57.0%	66.2%	-	100.0%
Portuguese	1.3%	9.5%	12.4%	3.4%	23.8%	-
Chinese	1.2%	2.3%	3.0%	.9%	5.8%	-
Creole (Haitian)	.6%	6.5%	6.3%	6.8%	16.2%	-
Vietnamese	.7%	2.1%	2.2%	1.9%	5.2%	-
Crioulo	.4%	5.7%	5.4%	6.3%	14.2%	-
Arabic	.4%	2.8%	3.0%	2.5%	7.0%	-
Russian	.3%	.5%	.7%	.2%	1.3%	-
Other language	.3%	1.2%	1.1%	1.4%	3.0%	-
French	.2%	1.3%	1.3%	1.5%	3.3%	-
Khmer	.2%	.7%	.4%	1.3%	1.7%	-

For the biology assessment sample, demographic characteristics for EBs, EB subsamples, and EPs are presented in Table 3.4. There did not appear to be any differences in gender based on subsample. Subgroups follow similar racial distributions to the mathematics assessment sample, but Asian students have a smaller presence in the non-Spanish-speaking EB sample. As a group, EBs have spent about half as much time as EPs enrolled in Massachusetts schools and in the same school district, but differences emerge when looking at length of time as an EB. LTEBs spend roughly as much time enrolled as EPs, with STEBs enrolled for a much shorter time. However, these enrollment differences between LTEBs and STEBs are an artifact of the way LTEB and STEB status was determined. Spanish-speaking EBs appear to be enrolled slightly longer than non-Spanish-speaking EBs. Similar trends to students in the mathematics assessment sample were observed for students identified as economically disadvantaged or homeless. EBs are identified as economically disadvantaged at a rate more than double than that of EPs, however a greater percentage of EPs have an individualized education plan (IEP) than EBs. When examining EB subgroups, 47.7% of LTEBs have an IEP compared to 7.0% of STEBs, and 22.1% of Spanish-speaking EBs have an IEP compared to 10.4% of non-Spanish-speaking EBs. EBs were identified as homeless at a much greater rate than EPs, but there were differences in homeless rates by EP sample. LTEBs and Spanish-speaking EBs appear to experience more homelessness than STEBs and non-Spanish-speaking EBs. The distribution of performance level on the biology assessment by subsample is similar to the distribution for the mathematics assessment, with 20.8% of non-Spanish-speaking EBs scoring proficient or advanced compared to 8.9% of Spanish-speaking EBs, 14.2% of STEBs, and 10.3% of LTEBs.

Table 3.4.

Demographic characteristics for students by subsample – Biology assessment

Characteristic	EP	EB	STEB	LTEB	OTH	SPA
n	15214	1922	1419	504	723	1199
Female	47.8%	47.0%	47.3%	46.1%	47.3%	46.8%
Male	52.1%	53.0%	52.6%	53.9%	52.7%	53.1%
Asian	6.9%	6.3%	6.8%	5.0%	16.9%	.0%
African-American/Black	9.2%	15.9%	15.4%	17.1%	41.9%	.2%
Hispanic or Latino	16.8%	67.8%	66.8%	70.8%	16.9%	98.6%
Multiracial, non-Hispanic or Latino	3.4%	.4%	.1%	1.2%	1.0%	.1%
American Indian or Alaskan Native	.3%	.2%	.2%	.0%	.0%	.3%
Native Hawaiian or Pacific Islander	.1%	.1%	.1%	.2%	.3%	.0%
White	63.3%	9.3%	10.5%	5.8%	23.1%	0.9%
Avg. years student attended MA schools	9.8	4.6	2.8	9.5	4.2	4.8
Avg. years student continuously enrolled in district	5.7	3.4	2.4	6.3	3.2	3.6
Economically Disadvantaged	33.8%	76.4%	75.8%	78.1%	70.5%	80.0%
IEP Status	23.9%	17.7%	7.0%	47.7%	10.4%	22.1%
Homeless	1.2%	10.1%	11.8%	5.6%	4.3%	13.7%
Advanced	24.2%	1.7%	2.1%	.4%	3.6%	.5%
Proficient	44.5%	11.6%	12.1%	9.9%	17.2%	8.2%
Needs improvement	21.4%	31.7%	30.4%	35.6%	36.1%	29.1%
Failing	9.8%	55.0%	55.4%	54.1%	43.2%	62.2%

In Table 3.5, the twelve most common first languages for students in the biology assessment sample are presented by subsample, with the most common first languages at the top and less common first languages at the bottom. To highlight how many different languages EBs speak, the percentage of students who speak one of these twelve languages was calculated ("% in a 'dominant' language category"). The "Other language" category represents those students who speak a language that is not in the DESE first language codes. The distributions of first languages between LTEBs and STEBs are similar, although there appear to be higher rates of Spanish and Crioulo-speaking LTEBs than STEBs and Portuguese and Arabic-speaking STEBs than LTEBs.

Table 3.5.

First languages for students by subsample – Biology assessment

First language	EP	EB	STEB	LTEB	OTH	SPA
n	15214	1922	1419	504	723	1199
English	87.6%	-	-	-	-	-
Spanish	5.0%	62.4%	60.0%	69.2%	-	100.0%
Portuguese	1.8%	11.2%	13.8%	4.0%	29.9%	-
Chinese	1.0%	2.4%	2.8%	1.2%	6.4%	-
Creole (Haitian)	.6%	5.0%	4.7%	6.2%	13.4%	-
Vietnamese	.3%	.8%	1.0%	.4%	2.2%	-
Crioulo	.9%	6.6%	5.6%	9.1%	17.4%	-
Arabic	.3%	2.5%	2.9%	1.6%	6.8%	-
Russian	.3%	.3%	.4%	.0%	.7%	-
Other language	.2%	.9%	1.1%	.6%	2.5%	-
French	.2%	1.0%	1.0%	1.0%	2.6%	-
Khmer	.3%	1.0%	.5%	2.4%	2.6%	-

Materials

As discussed previously data was drawn from two large-scale assessments: the 2019 10th grade mathematics MCAS and the 2019 high school biology MCAS (DESE, 2019a; DESE 2019b). Appendix D contains information on the features of the MCAS assessments used in the present study. Tables D1 and D2 present the item score descriptive statistics for the mathematics and biology assessments, respectively. Tables D3 and D4 present the item type, points possible and reporting categories for the mathematics and biology assessments, respectively. Both assessments addressed state standards for their respective subjects: the mathematics domains assessed were Algebra and Functions, Geometry, Number and Quantity, and Statistics and Probability, and the biology domains assessed were Anatomy and Physiology, Biochemistry and Cell Biology, Ecology, Evolution and Biodiversity, and Genetics. The mathematics assessment has 42 items (32 dichotomous and ten polytomous), and the biology assessment has 45 items (40 dichotomous and five polytomous). Both assessments are made of multiple choice ("selected response" on the mathematics assessment; items in this category with multiple points possible had multi-part items) and constructed response ("short answer" and "constructed response" on the mathematics assessments). On the mathematics assessment, 36 items were selected response (31 dichotomous and five polytomous) three items were short answer (all polytomous), four items were constructed response (all polytomous). On the biology assessment, all 40 dichotomous items were multiple choice, and all 5 polytomous items were constructed response. Tables D5 and D6 present the comparison group by item score correlations for the mathematics and biology assessments, respectively.

The LC of items was determined in Study One; the factor scores from the confirmatory factor analysis of linguistic features in Study One are used in the present study. In Study One, data was collected from a rubric adapted from Abedi et al. (2010) to measure the lexical and grammatical complexity in assessment items on two mathematics assessments and two biology assessments. Lexical features were measured by having two raters count the number of words in each item ("total words," using Microsoft Word), the unique general academic vocabulary in an item ("general academic vocabulary," or words that are uncommon when compared to a corpus and are unrelated to the construct measured on the assessment), and the number of words with seven or more letters. Grammatical features were measured by having four trained graduate students count the number of instances of particular grammar features not containing construct-relevant vocabulary. More details can be found in Chapter Two.

Separate multidimensional models were created for each subject, one for mathematics and one for biology. However, it was found that while there was evidence for a multidimensional model of LC higher-order factors of LC and grammatical complexity would not improve model fit, for both subjects. The counts of grammatical features using these rubric were transformed into factor scores for complex noun phrases and relative clauses; other grammatical features were not counted consistently enough to be included in the present study. Factor scores for lexical complexity using factor loadings for total words, general academic vocabulary, and number of words with seven or more letters were obtained from Study One as well.

The factor scores from these models are used as predictors of LC in the present study, which further examines one mathematics and one biology assessment used in Study One. Tables D7 and D8 present the lexical complexity, complex noun phrases, and relative clauses factor scores for each item for the mathematics and biology assessments used in the present study, respectively. These factor scores were derived from models that modeled the lexical complexity, complex noun phrases, and relative clauses factors with a mean of zero and variance of one; descriptive statistics for the factor scores used in the present study are in Table 3.6. While lexical complexity follows a fairly normal distribution, the grammatical features (complex noun phrases and relative clauses) are positively skewed, with most values falling below the mean. It was common for raters in Study One to count no complex noun phrases or relative clauses in an item. Counts of zero complex noun phrases (for all raters) are represented by a factor score of -.664 on the mathematics assessment and -.739 on the biology assessment. Counts of zero relative clauses (for all raters) are represented by a factor score of -.489 on the mathematics assessment and -.380

on the biology assessment. The relationship between LC and item difficulty was evaluated with EIRMs using HLM software (De Boeck & Wilson, 2004).

Table 3.6.

Descriptive Statistics for Linguistic Complexity Factor Scores in Present Study

Subject	Statistic	Lexical Complexity	Complex Noun Phrases	Relative Clauses
Mathematics	Mean	0.137	0.001	-0.054
	Standard Deviation	1.064	1.116	1.037
	Skewness	1.127	2.423	2.415
	Kurtosis	1.120	5.760	4.591
Biology	Mean	-0.093	0.031	-0.035
	Standard Deviation	1.136	0.929	0.765
	Skewness	0.321	2.057	2.436
	Kurtosis	-1.006	4.203	5.237

Procedure

To answer these hypotheses, the present study used a Rasch hierarchical generalized linear modeling (HGLM) framework for each assessment with item-level LC covariates presented in Equation 3.7, with one model for no predictors of LC, lexical complexity factor as a predictor, complex noun phrase factor as a predictor, relative clause factor as a predictor, and combinations of lexical complexity, complex noun phrase, and relative clause factor scores as predictors as appropriate. If a LC predictor model significantly improved model fit, that LC factor was included in a model with other LC factors that significantly improved model fit. These models were estimated with version 7 of the HLM software program which utilizes penalized quasi-likelihood for HGLMs (Raudenbush et al., 2011). The item-level data was prepared in long format, with each row representing an examinee's response to an item; each row contained item indicator variables, with an item indicator of "1" indicating the examinee's response was to that item, with all other item indicator variables set to “0”. To evaluate the directionality of item estimates from this item indicator coding (i.e., do larger item logits correspond to easier or more difficult items?), dichotomous item estimates from the model evaluating DIF between EPs and EBs in the biology assessment were correlated with the percent of examinees responded correctly to each dichotomous item. A strong positive correlation indicates larger item logits correspond to easier items and a strong negative correlation indicates larger item logits correspond to more difficult items. The resulting correlation, r = -.996, indicated positive item logits corresponded to more difficult items and negative item logits corresponded to easier items.

Although the 2019 MCAS assessments were calibrated using a combination of a three-parameter logistic model (DESE, 2020a; DESE, 2020b), a two-parameter logistic model, and a graded-response model, the present study used a Rasch modeling approach (a rating scale model as a hierarchical model) due to computational constraints. Due to the large number of test-takers and parameters that need to be estimated, EIRMs may not converge if item discrimination and guessing parameters are accounted for.

Although DESE releases the item-level data for all students responses, only the responses of students who received an assessment score (e.g. “Advanced,” “Proficient,” “Needs Improvement,” “Failing”) were included. These students had sufficient assessment data (and in the case of EBs, were enrolled in Massachusetts schoolers longer than a year) to be given a score by DESE on the assessment, thus the results of this study can be generalized to Massachusetts students taking these assessments that received a score on these assessments. There are three key IRT assumptions: unidimensionality of a single latent person ability, local independence of item responses, and person responses to items can be modeled by an ogive curve (de Ayala, 2022). Unidimensionality can be evaluated with a parallel analysis (O’Connor, 2000). Results revealed two factors should be extracted for the mathematics assessment and two factors should be extracted for the biology assessment, suggesting that some other latent ability is measured by these assessments, although de Ayala notes there is usually some degree of violation for the unidimensionality assumption. The local independence assumption is concerned with a person's response to an item is a result of their underlying latent ability on the measured construct and not other latent abilities; the present study posits that the linguistic complexity in items is negatively influencing the responses of one group of test-takers (EBs) and not another group (EPs). If DIF is found in items on these assessments, then the local independence assumption may be violated; routinely screening items for DIF can evaluate whether this assumption is met.

Anchor Item Selection

In HGLMs, DIF estimates are calculated in reference to the reference item. This means that if the reference item selected is biased, this bias will influence each item's DIF estimate. Chen et al. (2014) illustrated this effect by describing that when an item with significant DIF is selected as the reference item, the DIF estimates of all the other items are shifted and may incorrectly flag items as exhibiting significant DIF which increases Type I error. Because of the influence of the reference item, an anchor item strategy needs to be applied to select a reference item that is bias-free to obtain accurate DIF estimates. However, this effect can also be mitigated by evaluating DIF estimates to include the effect of the reference item.

Anchor items are those items that are presumed to be DIF-free (Kopf et al. 2015). Anchor items must be determined prior to DIF analyses. It is vital to select a first anchor item that is DIF-free, otherwise we may not have accurate results for whether there are group differences in item responses (Kopf et al., 2013). To identify anchor items in HGLMs, the constant item (CI) method can be used; the CI method has been found to do well controlled Type I error even on assessments with a high percentage of DIF items (Chen et al., 2014, Shih & Wang, 2009). With the CI method, the following steps are implemented:

1. Using the model in Equation 3.4, set one item as the reference item and evaluate all other items for DIF, constraining mean ability between groups to zero (Shih & Wang, 2009). A DIF estimate (γ_{q1} for item q) is obtained for each item but the reference item.
2. Step one is followed k times for k items, with each item set as the reference item and all other items are assessed for DIF, with DIF estimates for each item.
3. Calculate the mean absolute values of each item's DIF estimates across k -1 models.
4. The item with the smallest mean absolute value DIF estimate is selected as the anchor item.

Chen et al. (2014) found this method to have satisfactory power in controlling Type I error rates and also found using one anchor item controlled Type I error rate better than using four anchor items due to the lower probability of including items with DIF in the anchor set. Although anchor selection with non-HGLM methods using an iterative purification procedure favor increasing the number of anchor items (Kopf et al., 2015), this method may not be appropriate for HGLMs given the results of Chen et al. (2014). Other researchers have found anchor sets with one item have comparable rates of power and Type 1 error rates to anchor sets with more items when sample sizes are large (n > 1,000), as they are in the present study (Shih & Wang, 2009). Although the item with the smallest mean absolute DIF estimates may not truly be free, by selecting the item with the least amount of DIF, the Type I error rate can be reduced, leading to less biased DIF identification; including items with DIF in the anchor set increases Type I error rate.

When conducting multiple analyses with different sets of comparison groups, to compare results for different groups, the same first anchor item should be used across groups. However, when conducting multiple analyses with different sets of comparison groups, to compare results for different groups, the same first anchor item should be used across groups. The CI method results suggested different items for the first anchor item for most comparison groups. To resolve this discrepancy in selecting the first anchor item, results from the CI method analyses were compared across comparison groups (see Appendix D for the complete comparative table of results). There were many items with similar, low values for total DIF effect; it was determined that the anchor item should be an item with similar low total DIF effect across comparison groups. If an item has low total DIF effects across groups, then it is likely to be DIF-free, or at least DIF-free between students grouped by differing levels of English proficiency. For example, although m07, m25, m31, and m42 exhibited the lowest DIF in the EPvEB mathematics assessment anchor item analysis, these items did not exhibit the lowest DIF in the EP versus EB subgroup comparisons. Therefore, the items that consistently demonstrated low mean DIF effect across comparison groups' anchor item analyses were chosen as the anchor item for each subject; this item was also set as the reference item to simplify interpretations.

For the mathematics assessment, Item m13 was selected as the first anchor item for HGLM analyses due to m13's consistently low mean absolute value DIF effect across comparison groups. Table 3.7 presents a summary of anchor item selection results based on the CI method for the mathematics assessment across comparison groups. Many items' total DIF effect for each comparison groups' CI method analysis were close to the item with lowest mean DIF effect. While m13 was not the item with the lowest mean absolute value DIF effect, m13's mean absolute value DIF effect was close to the item with the lowest mean absolute value DIF effect for all comparison groups except for LTEBvSTEB; m13 also had the lowest mean absolute value DIF effect when the mean absolute value DIF effects were averaged across comparison groups. The LTEBvSTEB and OTHvSPA comparison groups had the lowest amounts of DIF present in the CI method analyses; DIF was often not statistically significant for these groups. As these comparison groups are comprised of only EBs, large effects of DIF between EBs is not expected due to similar levels of English proficiency in test-takers. Therefore, selecting m13 as the anchor item for the LTEBvSTEB set of HGLM analyses should not unduly influence the effect of DIF since there is not much DIF to be found between LTEBs and STEBs. In addition, m13 had the lowest average mean absolute value DIF effect across comparison groups. This item was also set as the reference item.

Table 3.7.

Anchor Item Selection Results Summary – Mathematics Assessment

Comparison Group	Lowest Mean DIF Effect	Item	m13 Mean DIF Effect	m13 Ranking for Lowest Mean DIF Effect	Mean Item Mean DIF Effect
EBvEP	0.88	m41	0.89	7th	1.34
EPvSTEB	0.90	m41	0.91	10th	1.35
EPvLTEB	0.89	m36	0.89	5th	1.34
LTEBvSTEB	0.14	m40	0.22	30th	0.20
ENGvSPA	0.95	m08	0.97	10th	1.45
ENGvOTH	0.80	m41	0.80	6th	1.21
OTHvSPA	0.23	m36	0.23	6th	0.33

Item b02 was selected as the anchor item for HGLM analyses due to b02's consistently low mean absolute value DIF effect across comparison groups. Table 3.8 presents a summary of anchor item selection results based on the CI method for the biology assessment across comparison groups. Many items' mean absolute value DIF effect for each comparison groups' CI method analysis were close to the item with lowest mean absolute value DIF effect. While b02 was not the item with the lowest mean absolute value DIF effect, b02's mean absolute value DIF effect was close to the item with the lowest mean absolute value DIF effect for all comparison groups; b02 also had the lowest mean absolute value DIF effect when the mean DIF effects were averaged across comparison groups. As for the mathematics assessment, the LTEBvSTEB and OTHvSPA comparison groups had the lowest amounts of DIF present in CI method analyses; DIF was often not statistically significant for these groups. As these comparison groups are comprised of only EBs, large effects of DIF between EBs is not expected due to similar levels of English proficiency in test-takers. In addition, b02 had the lowest average mean absolute value DIF effect across comparison groups. This item was also set as the reference item.

Table 3.8.

Anchor Item Selection Results Summary – Biology Assessment

Comparison Group	Lowest Mean DIF Effect	Item	b02 Mean DIF Effect	b02 Ranking for Lowest Mean DIF Effect	Mean Item Mean DIF Effect
EBvEP	0.67	b38	0.67	2nd	1.05
EPvSTEB	0.68	b04	0.68	3rd	1.07
EPvLTEB	0.66	b31	0.66	6th	1.03
LTEBvSTEB	0.18	b18	0.19	13th	0.27
ENGvSPA	0.72	b31	0.72	2nd	1.14
ENGvOTH	0.60	b38	0.60	6th	0.93
OTHvSPA	0.19	b16	0.19	8th	0.29

Model Building

To evaluate whether the inclusion of an LC item covariate influences DIF between multiple comparison groups (see Table 3.1 for coding of comparison groups), with each comparison group comprising its own dataset. Several models were created; in each model a different item served as the reference item while all other items were assessed for DIF. The following steps were followed for each comparison group for each assessment after the reference item was established with the CI method. First a “comparison model” was created examining only main effect of group with no items assessed for DIF, for the purpose of examining model fit compared to a “base model” that examined DIF between the reference and focal groups for all items but the reference item (Equation 3.8, the same model as Equation 3.7 but without any item characteristic s).

$$
\begin{aligned}
\eta_{0ij} &= \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} \\
\eta_{1ij} &= \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} + \delta_1 \\
\eta_{2ij} &= \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} + \delta_2 \\
\eta_{3ij} &= \beta_{0j} + \sum_{q=1}^{k-1} \beta_{qj} X_{qij} + \delta_3 \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(G_j) + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}(G_j) \\
&\vdots \\
\beta_{(k-1)j} &= \gamma_{(k-1)0} + \gamma_{(k-1)1}(G_j) \\
&\delta_1 \\
&\delta_2 \\
&\delta_3
\end{aligned}
\tag{3.8}
$$

The lexical complexity ("LEX predictor"), complex noun phrases ("NP predictor"), and relative clauses ("RC predictor") factor scores were added to the base model as three separate models as LC item covariates; the interaction of the LC item covariate with focal group status was also evaluated (Equation 3.9, the same model as Equation 3.7, but with a LC predictor as item characteristic *s*).

$$\eta_{0ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj}X_{qij}$$

$$\eta_{1ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj}X_{qij} + \delta_1$$

$$\eta_{2ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj}X_{qij} + \delta_2$$

$$\eta_{3ij} = \beta_{0j} + \beta_{sj}(Y_{sqi}) + \sum_{q=1}^{k-1} \beta_{qj}X_{qij} + \delta_3$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(G_j) + u_{0j} \tag{3.9}$$

$$\beta_{sj} = \gamma_{s0} + \gamma_{s1}(G_j) + u_{sj}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(G_j)$$

$$\vdots$$

$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(G_j)$$

$$\delta_1$$

$$\delta_2$$

$$\delta_3$$

Models with multiple LC predictors were created if the models with a single LC predictor significantly improved model fit. Omnibus tests (likelihood ratio tests) were conducted between the comparison model and the base model to evaluate if including focal group status and evaluating DIF in items improved model fit. Even if the base model does not significantly improve model fit, DIF was still evaluated in items as the overarching goal of the present dissertation was to explore the effects of LC on item responses for how these linguistic features affect all test-takers and potentially explain group differences in item responses between EBs and non-EBs. Omnibus tests were conducted between the base model and LC predictor models to evaluate whether inclusion of an LC factor predictor and its interaction with focal group status led to increased model fit. If single LC predictor models improved model fit, models with multiple LC predictors were created to see if the inclusion of multiple LC predictors improved model fit compared to single LC predictor models. The omnibus test is an overall test of fit using a likelihood ratio test, if $p < .01$ for the omnibus tests, model fit improved. Other measures of model fit were examined. The AIC and BIC of the models were compared; models with lower AIC or BIC values suggested improved model fit.

To determine which group had higher ability estimates before and after conditioning for LC predictors (if LC predictors improved model fit), the significance and direction of the intercept's interaction with focal group status (γ_{01}) was examined for each model. To answer Hypothesis 1 ("LC factor scores will have significant main effects and interactions with emergent bilingual status; the interactions will favor English proficient students"), the significance of LC predictors' main effects (γ_{s0}) and their interactions with focal group status (γ_{s1}) were examined. If a model with an LC item covariate has a significant main effect ($p <$ .05), then that predictor influences item responses. For positive LC predictor main effects, items with a higher LC predictor factor score are associated with increased ability estimates than items with a lower LC predictor factor score; for negative LC predictor main effects, items with a higher LC predictor factor score are associated with decreased ability estimates than items with a lower LC predictor factor score. If a model with an LC item covariate has a significant interaction with focal group status (γ_{s1}, $p < .05$), then that predictor has an effect on group differences in item responses, or DIF. To answer Hypotheses 2 and 3 ("For items with higher LC, there will be less items flagged as significantly favoring EPs when including LC as a covariate" and "For items with lower LC, there will be no change in items flagged as significantly favoring EPs when including LC as a covariate") DIF results between the base model and LC factor predictor models were compared to evaluate which items changed DIF significance or direction upon the inclusion of an LC factor. Positive values of γ_{q1} indicate DIF favoring the focal group and negative values of γ_{q1} indicate DIF favoring the reference group. The LC factor scores of items with DIF were examined and rated high, medium, or low based on the factor scores relation to the median factor score of that LC feature. High LC factor scores were at least one standard deviation above the mean, medium LC factor scores were around the median value, and low LC factor scores were those with the lowest values; due to the positive skew of LC factor scores low LC factor scores were about half a standard deviation below the mean, therefore only the lowest LC factor scores were examined for "low" LC. For complex noun phrases and relative clauses, low factor scores indicated that feature was not present in the item. Even if DIF is not detected in the base model (as may be the case for EB versus EB comparison groups), the effects of LC predictors will still be examined because identifying significance of the main effects of LC and LC's interactions with focal group status will provide insight into how LC influences the item responses of test-takers.

While significant item by focal group status interactions (γ_{q1}) indicate significant DIF, the practical significance of the DIF identified must be considered, especially with a sample size this large. In preliminary analyses using a $p < .05$ cut-off, many items were identified as having significant DIF. However, γ_{q1} only indicates DIF in reference to the reference item. To get a true estimate of DIF, the significance of the combined term of $\gamma_{q1} + \gamma_{01}$ must be evaluated, which makes traditional significance testing more difficult. To evaluate the significance of this combined term for each item, 95% confidence intervals were created for each adjusted DIF estimate using the standard error from the original DIF estimate. If the confidence interval contained the value for γ_{01}, the item did not exhibit significant DIF. To ensure the identification of sizable group differences on item responses, ETS's procedure for classifying the magnitude of DIF was utilized (Zwick, 2012; Monahan, et al., 2007). By taking the odds-ratios of the item by focal group status interaction plus the group differences in item responses ($\gamma_{q1} + \gamma_{01}$) and using Equation 3.8, the magnitude of DIF can be interpreted for the base model (Monahan, et al., 2007). For the models including LC predictors, the item by focal group status interaction, group differences in item responses, and LC predictor by focal group interaction were summed together ($\gamma_{q1} + \gamma_{01} + \gamma_{s1}$) to calculate the adjusted DIF estimates. Like in the base model, 95% confidence intervals were created to evaluate the significance of the adjusted DIF estimates and odds-ratios were used to determine the effect size of DIF. As the focal group effect of the LC predictor on the item has been partitioned out of the DIF estimate, it needs to be added back to the DIF estimate for that item to determine the total effect size of DIF.

$$\Delta OR = -2.35 ln(OR) \tag{3.8}$$

Zwick (2012) discussed ETS's classification rules. When $|\Delta OR|$ < 1.00, there is negligible DIF ("Category A"); when $|\Delta OR|$ > 1.00 or < 1.50, there is moderate DIF ("Category B"); when $|\Delta OR|$ > 1.50, there is substantial DIF ("Category C"). Liu & Bradley (2021) compared using this method for an HGLM DIF model and compared it to a traditional Mantel-Haenszel procedure and a Rasch analysis in WINSTEPs. The same items were flagged for DIF between all three methods, with the Mantel-Haenszel and WINSTEPs procedures flagging more and different items. This suggests this HGLM DIF effect size method is more conservative in detecting noticeable DIF.

Results

The results for MCAS Mathematics are discussed before MCAS Biology. Summaries of results across subgroups are presented before more specific results by subgroup. Each analysis for each subject will be presented in the following order (see Table 3.7 for code references): EPvEB, EPvSTEB, EPvLTEB, STEBvLTEB, EPvSPA, EPvOTH, OTHvSPA. The codes also denote that the first group listed is the reference group (value of "0") and the second group listed is the focal group (value of "1"). First, omnibus test results and AIC and BIC values of models are presented to determine if the base model (examining DIF in all items but the reference item) fits better than the comparison model (focal group status and DIF not examined), if models including LC predictors fit better than the base model, and if models with multiple LC predictors fit better than models with a single LC predictor.

After determining what linguistic features improve the base model fit, the significance and direction of the intercept's interaction with focal group status (γ_{01}) was examined for each model to determine which group had higher ability estimates before and after conditioning for LC predictors. Next, the significance of LC predictors' main effects (γ_{s0}) and their interactions with focal group status (γ_{s1}) were examined. DIF results between the base model and LC factor predictor models were compared to evaluate which items changed DIF significance or direction after conditioning for an LC predictor.

Next, each comparison group is discussed, and the specific omnibus test results and the base model's item difficulties for the reference and focal groups are presented for each comparison group. Polytomous item thresholds have decimal points in their labels to denote the number of points given for that threshold (e.g., m09.1 indicates the threshold to score at least one point on item m09). Tables with item difficulties for each comparison group (for the sample of test-takers in the comparison group, focal group, reference group, and the differences between the focal and reference groups' item difficulties) are located in Appendix E. Model results are then discussed (see Appendix F for tables of model results); tables are presented for each comparison groups' model results which include estimates, standard errors, and *p*-values for each estimated parameter, along with effect sizes for determining the magnitude of DIF. Tables are also presented for the calculation of the adjusted DIF estimates and 95% confidence intervals used to determine the statistical significance of the adjusted DIF estimates. A hyphen denotes that parameter was not included in the model, such as the reference item (m13 for the mathematics assessment and b02 for the biology assessment) and the reference item's interaction with EB status.

For the EPvEB comparison group, the items changing DIF significance or direction are discussed. The magnitude of DIF was determined by using ETS's classification rules for substantial or moderate DIF, as discussed in the methods section. For the EP versus EB subgroup comparison groups, the changes between these comparison groups are the EPvEB comparison group are discussed. For the EB versus EB comparison groups, items changing DIF significance or direction from the base model to LC predictor models are discussed.

For each comparison group and LC factor, the significance of the LC factor main effect and interaction with focal group status are discussed to evaluate Hypothesis 1: "LC factor scores will have significant main effects and interactions with emergent bilingual status; the interactions will favor English proficient students." For each comparison group and LC factor, the changes in DIF significance or direction between the base model and an LC factor predictor model were examined to answer Hypotheses 2: "For items with higher LC, there will be less items flagged as significantly favoring EPs when including LC as a covariate" and 3: "For items with lower LC, there will be no change in items flagged as significantly favoring EPs when including LC as a covariate," the significance of the item by EB status interactions between models will be compared. If an item by EB status interaction (test of item's DIF) went from being significant in the base model to non-significant in a model using a linguistic feature as a predictor, then that linguistic feature in that item may explain that item's DIF in the base model if the item's LC factor score by EB status interaction is also significant.

MCAS Mathematics Results

Summary of Subgroup Results for the Mathematics Assessment

Due to conducting many HGLMs, summaries of the results are presented and discussed first, starting with model fit comparisons in Table 3.9. Significant changes in log likelihood and lower AIC and BIC values indicated improvements in model fit. For all models, when changes in log likelihood was significant, AIC and BIC were lower, and when changes in log likelihood were significant, AIC and BIC were higher, therefore model fit results agreed with each other. The results of the omnibus tests and AIC and BIC comparisons demonstrate the base model assessing DIF did not improve model fit compared to the comparison model that did not assess DIF. However, this does not mean DIF should not be investigated further, as identifying DIF is a crucial step in the test development process to ensure valid interpretations of scores for all test-takers. For the EP versus EB comparison groups, neither the LEX predictor nor RC predictor models improved model fit compared to the base model. However, the NP predictor model improved model fit compared to base model for all EP versus EB comparison groups. As only the NP predictor model improved model fit, models with multiple LC predictors were not examined for the EP versus EB comparison groups.

For the EB versus EB comparison groups, each LC predictor model improved model fit compared to both the comparison and base models. As the LEX predictor and RC predictor models improved fit the most, a multiple LC predictor model with both of these factors was examined next, this model improved fit both EB versus EB comparison groups. A model with all three LC predictors improved model fit compared to the model with the LEX and RC predictors, suggesting that the inclusion of these LC factor scores explains EBs' item responses on this mathematics assessment. Specifics of the omnibus tests and changes in AIC and BIC for the inclusion of each LC factor are presented in the results for each comparison group.

Table 3.9.

Summary of Model Fit Improvement for each Comparison Group – Mathematics Assessment

Comparison Group	Improved Model Fit Compared to:	Base	LEX	NP	RC	LEX + RC	All predictors
EPvEB	Comparison	X	-	-	-	-	-
	Base	-	X	✓	X	-	-
	Any Single LC Predictor	-	-	-	-	-	-
	LEX + RC	-	-	-	-	-	-
EPvSTEB	Comparison	X	-	-	-	-	-
	Base	-	X	✓	X	-	-
	Any Single LC Predictor	-	-	-	-	-	-
	LEX + RC	-	-	-	-	-	-
EPvLTEB	Comparison	X	-	-	-	-	-
	Base	-	X	✓	X	-	-
	Any Single LC Predictor	-	-	-	-	-	-
	LEX + RC	-	-	-	-	-	-
STEBvLTEB	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
EPvSPA	Comparison	X	-	-	-	-	-
	Base	-	X	✓	X	-	-
	Any Single LC Predictor	-	-	-	-	-	-
	LEX + RC	-	-	-	-	-	-
EPvOTH	Comparison	X	-	-	-	-	-
	Base	-	X	✓	X	-	-
	Any Single LC Predictor	-	-	-	-	-	-
	LEX + RC	-	-	-	-	-	-
OTHvSPA	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓

Note: "✓" indicates significant changes in -2 log likelihood and lower AIC and BIC values that led to a judgement of improved model fit. "X" indicates non-significance.

Table 3.10 presents the significance of the intercept's interaction with focal group status (γ_{01}) for each comparison group for the mathematics assessment. Positive interactions indicated the focal group had higher ability estimates; negative interactions indicated the reference group had higher ability estimates. Readers should note the base model examines DIF, but does not include any LC predictors, the LEX predictor model only includes the lexical complexity predictor, the NP predictor model only includes the complex noun phrases predictor, the RC predictor model only includes the relative clauses predictor, and the all predictors model includes all three LC predictors. For the EP versus EB comparison groups, EPs consistently had higher ability estimates in the base model, but when complex noun phrases and its interaction with focal group status were included in the model, EBs had higher ability estimates. For the EB versus EB comparison groups, there were no significant group differences in ability estimates for STEBvLTEB in the base model, but non-Spanish-speakers had higher ability estimates than Spanish-speakers in the OTHvSPA base model. When only lexical complexity and its interaction with focal group status were accounted for in the EB versus EB comparison group models, LTEBs and Spanish-speaking EBs had higher ability estimates. When only complex noun phrases and its interaction with focal group status were accounted for in the EB versus EB comparison group models, there were no significant group differences in ability estimates. When only relative clauses and its interaction with focal group status were accounted for in the EB versus EB comparison group models, LTEBs and non-Spanish-speaking EBs had higher ability estimates. When all LC predictors and their interaction with focal group status were accounted for in the EB versus EB comparison group models, there were no significant group differences in ability estimates.

Table 3.10.

Significance of the Intercept's Interaction with Focal Group Status (γ_{01}) for each Comparison Group – Mathematics Assessment

Comparison Group	Base Model	LEX	NP	RC	All Predictors
EPvEB	Favors EPs ***	-	Favors EBs ***	-	-
EPvSTEB	Favors EPs ***	-	Favors STEBs ***	-	-
EPvLTEB	Favors EPs ***	-	Favors LTEBs ***	-	-
STEBvLTEB	0.351	Favors LTEBs ***	0.112	Favors LTEBs ***	0.254
EPvSPA	Favors EPs ***	-	Favors SPAs ***	-	-
EPvOTH	Favors EPs ***	-	Favors OTHs ***	-	-
OTHvSPA	Favors OTHs ***	Favors SPAs ***	0.254	Favors OTHs **	0.396

Note: *** = p < .001, ** = p < .01. If γ_{01} was not significant, p-values were listed instead.

Table 3.11 presents the significance of LC Factor predictors' main effects (γ_{s0}) and their interactions with focal group status (γ_{s1}) for each comparison group for the mathematics assessment for the models with a single LC predictor. Negative interactions indicated items with higher LC factor scores were easier for the focal group; positive interactions indicated items with higher LC factor scores were easier for the reference group. For all comparison groups, the significant positive main effect of complex noun phrases indicated that items with higher complex noun phrases factor scores were associated with increased ability estimates than items with lower complex noun phrases factor scores. For the EB versus EB comparison groups, the significant positive main effects of lexical complexity indicated that items with higher lexical complexity factor scores are associated with increased ability estimates than items with lower lexical complexity factor scores, and the significant positive main effects of relative clauses indicated that items with higher relative clauses factor scores were are associated with increased ability estimates than items with lower relative clauses factor scores.

For the EP versus EB comparison groups, items with higher complex noun phrases factor scores were consistently easier for EPs than for EBs. However, differences between the reference and focal groups emerged when examining the EB versus EB comparison groups. In the LEX predictor model, items with higher lexical complexity factor scores were easier for STEBs and non-Spanish-speaking EBs than LTEBs and Spanish-speaking EBs, respectively. In the NP predictor model, there were no group differences in how complex noun phrases influenced item responses for the STEBvLTEB comparison group, but items with higher complex noun phrases factor scores were easier for non-Spanish-speaking EBs than for Spanish-speaking EBs. In the RC predictor model, items with higher relative clauses factor scores were easier for STEBs and non-Spanish-speaking EBs than LTEBs and Spanish-speaking EBs, respectively.

Table 3.11.

Significance of LC Factor Predictors (γ_{s0}) and Interactions with Focal Group Status (γ_{s1}) for each Comparison Group for Single LC Predictor Models – Mathematics Assessment

Comparison Group	LEX		NP		RC	
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**
EPvEB	-	-	***	Favors EPs ***		-
EPvSTEB	-	-	***	Favors EPs ***	-	-
EPvLTEB	-	-	***	Favors EPs ***	-	-
STEBvLTEB	***	Favors STEBs ***	*	0.144	***	Favors STEBs ***
EPvSPA	-	-	***	Favors EPs ***	-	-
EPvOTH	-	-	***	Favors EPs ***	-	-
OTHvSPA	***	Favors OTHs ***	*	Favors OTHs *	***	Favors OTHs ***

Note: *** = $p < .001$, * = $p < .05$. If γ_{s1} was not significant, *p*-values were listed instead.

Table 3.12 presents the significance of LC factor predictors' main effects and their interactions with focal group status for the EB versus EB comparison groups for the mathematics assessment when all LC predictors are included in the model. For STEBvLTEB, the significant positive main effect of lexical complexity indicates that items with higher lexical complexity factor scores are associated with increased ability estimates than items with lower lexical complexity factor scores, the negative main effect of complex noun phrases indicates that items with higher complex noun phrases factor scores are associated with decreased ability estimates than items with lower complex noun phrases factor scores, and the non-significant main effect of relative clauses indicates that relative clauses did not influence ability estimates. For OTHvSPA, the significant positive main effect of lexical complexity indicates that items with higher lexical complexity factor scores are associated with increased ability estimates than items with lower lexical complexity factor scores, the negative main effect of complex noun phrases indicates that items with higher complex noun phrases factor scores are associated with decreased ability estimates than items with lower complex noun phrases factor scores, and the significant positive main effect of relative clauses indicates that items with higher relative clauses factor scores are associated with increased ability estimates than items with lower relative clauses factor scores. When complex noun phrases is the only LC predictor, higher complex noun phrases factor scores indicate higher ability estimates, but when lexical complexity and relative clauses are accounted for in the all predictors models, lower complex noun phrases factor scores indicate higher ability estimates. There were no significant interactions between any LC predictor and focal group status for either EB versus EB comparison group.

Table 3.12.

Significance of LC Factor Predictors and Interactions with Focal Group Status for EP Versus EB Comparison Groups for All Predictor Models – Mathematics Assessment

Comparison Group	**LEX**		**NP**		**RC**	
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**
STEBvLTEB	***	0.615	***	0.892	0.347	0.789
OTHvSPA	***	0.314	***	0.108	*	0.082

Note: *** = $p < .001$, * = $p < .05$. If γ_{s1} was not significant, *p*-values were listed instead.

The number of items changing DIF significance or direction with the inclusion of an LC factor and its interaction with focal group status are presented in Table 3.13. This section provides an overview of the items changing DIF significance or direction with specific results presented for each comparison group below, which includes the differences between the EPvEB models and EP versus subgroups of EB models The magnitude of DIF can be interpreted by taking the odds-ratios of the item by focal group status interaction and using ETS's classification rules (Monahan, et al., 2007). In the present study's base model, the odds-ratio is taken of the sum of the item by focal group status interaction plus group differences in item responses ($\gamma_{q1} + \gamma_{01}$) to account for the effect of group differences in responding to the reference item. For the models including LC predictors, the odds-ratio of the sum of the item by focal group status interaction, group differences in item responses, and LC predictor by focal group interaction ($\gamma_{q1} + \gamma_{01} + \gamma_{s1}$) is used to determine the effect size of DIF after conditioning for LC predictors. When ΔOR > 1.50 and the adjusted DIF estimate's 95% confidence interval is above γ_{01}, there is substantial DIF favoring the reference group. When ΔOR < 1.50 and > 1.00 and the adjusted DIF estimate's 95% confidence interval is above γ_{01}, there is moderate DIF favoring the reference group. When ΔOR > 1.50 and the adjusted DIF estimate's 95% confidence interval is below γ_{01}, there is substantial DIF favoring the focal group. When ΔOR < 1.50 and > 1.00 and the adjusted DIF estimate's 95% confidence interval is below γ_{01}, there is moderate DIF favoring the focal group. If there is DIF in the base model and DIF is not present after conditioning for LC predictors, LC may be a source of bias in items. If there is DIF in the base model and DIF is present after conditioning for LC predictors and does not change which group is favored, there is no evidence for LC as a source of bias in items. If there is no DIF in the base model and DIF is present after conditioning for LC predictors, then there may be some other factor that is a source of bias in items that is mitigated by the effect of LC.

For the EPvEB comparison groups, many items that exhibited DIF favoring EBs in the base model were the polytomous items on the mathematics assessment. Examination of the thresholds revealed meeting these higher-point thresholds were either biased in favor of EPs or were non-significant. Therefore, these items that favored EBs in the base model were indicators that EBs had an easier time achieving the one-point threshold than EPs. When accounting for complex noun phrases factor scores, these items did not change DIF significance or direction. For the dichotomous items, items had a mix of DIF effects, with items exhibiting DIF favoring EPs, DIF favoring subgroups of EBs, or non-significant DIF. Accounting for complex noun phrases factor scores however led to the majority of these items favoring EBs and EB subgroups, although some items remained exhibiting non-significant DIF or switched from exhibiting significant DIF to exhibiting non-significant DIF. For the EBvEB comparison groups, items generally did not exhibit DIF in the base model or after accounting for an LC predictor, but some items favored the focal group after accounting for an LC predictor. For STEBvLTEB, all items exhibited non-significant DIF in the base model, and accounting for any or all LC predictor did not lead to changes in DIF significance. For OTHvSPA, most items exhibited non-significant DIF in the base model and accounting for any or all LC predictors generally did not lead to changes in DIF significance, but some items exhibited DIF favoring Spanish-speaking EBs in the base model or after accounting for an LC predictor, suggesting these LC features may interplay with one another for this comparison group.

Tables 3.14 and 3.15 show the items changing DIF significance or direction with the inclusion of an LC factor and its interaction with focal group status for items with high or low LC factor scores, respectively. For the all predictors models, only items with two or more high LC factors and no low LC factors (Table 3.14) or two or more low LC factors and no high LC factors (Table 3.15) were considered. Items were considered as having a high LC factor score when the LC factor score was greater than one standard deviation above the mean. Items were considered as having a low LC factor score when the LC factor score was greater than one standard deviation below the mean for lexical complexity or was the lowest LC factor score value for complex noun phrases and relative clauses.

For the EP versus EB comparison groups, the items with high complex noun phrases factors scores, m14, m33, m35, m38, and m40, generally exhibited substantial DIF favoring EBs before and after complex noun phrases were accounted for. The exception to this was m38 which exhibited moderate DIF favoring EPs in the base model and moderate DIF favoring EBs in the NP predictor model for EPvEB and EPvSPA, exhibited moderate DIF favoring EPs in the base model and non-significant DIF in the NP predictor model for EPvSTEB and EPvLTEB, and exhibited non-significant DIF in the base model and the NP predictor model for EPvOTH. For EPvOTH, three items with high complex noun phrases factor scores that exhibited DIF favoring EBs in the base model exhibited non-significant DIF after complex noun phrases were accounted for. For the EP versus EB comparison groups, the items with low complex noun phrases factor scores, m07, m09, m09, m11, m21, m24, m26, and m39, generally exhibited DIF favoring EPs after complex noun phrases were accounted for regardless of the direction or significance of DIF in the base model, except for EPvOTH, where more items exhibited non-significant DIF after accounting for complex noun phrases. For STEBvLTEB, all items exhibited non-significant DIF before and after accounting for any LC predictors. For OTHvSPA, few items exhibited significant DIF in the base model and most of the items that did exhibit significant DIF in the base model went from exhibiting DIF favoring Spanish-speaking EBs to exhibited non-significant DIF after accounting for any LC predictor. These items tended to be items with high or low LC factor scores, although some items exhibited DIF favoring Spanish-speaking EBs before and after accounting for an LC predictor. Specific details are discussed further in the subgroup comparison results.

Table 3.13.

Number of Items Changing DIF Significance or Direction After Conditioning on Linguistic Complexity – Mathematics Assessment

Analysis	Base Model DIF Direction	LEX			NP			RC			All Predictors		
		Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal
EPvEB	Favor Ref	-	-	-	0	0	14	-	-	-	-	-	-
	No DIF	-	-	-	0	6	6	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	16	-	-	-	-	-	-
EPvSTEB	Favor Ref	-	-	-	0	1	15	-	-	-	-	-	-
	No DIF	-	-	-	0	6	4	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	16	-	-	-	-	-	-
EPvLTEB	Favor Ref	-	-	-	0	1	12	-	-	-	-	-	-
	No DIF	-	-	-	0	4	9	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	16	-	-	-	-	-	-
STEBv LTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	42	0	0	42	0	0	42	0	0	42	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0
EPvSPA	Favor Ref	-	-	-	0	0	18	-	-	-	-	-	-
	No DIF	-	-	-	0	5	3	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	16	-	-	-	-	-	-
EPvOTH	Favor Ref	-	-	-	0	1	7	-	-	-	-	-	-
	No DIF	-	-	-	0	12	8	-	-	-	-	-	-
	Favor Focal	-	-	-	0	3	11	-	-	-	-	-	-
OTHvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	39	0	0	39	0	0	36	3	0	39	0
	Favor Focal	0	2	1	0	2	1	0	2	1	0	3	0

dd

Table 3.14.

Number of High LC Items Changing DIF Significance or Direction After Conditioning on Linguistic Complexity – Mathematics Assessment

Analysis	Base Model DIF Direction	LEX			NP			RC			All Predictors		
		Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal
EPvEB	Favor Ref	-	-	-	0	0	1	-	-	-	-	-	-
	No DIF	-	-	-	0	0	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	4	-	-	-	-	-	-
EPvSTEB	Favor Ref	-	-	-	0	1	0	-	-	-	-	-	-
	No DIF	-	-	-	0	0	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	4	-	-	-	-	-	-
EPvLTEB	Favor Ref	-	-	-	0	1	0	-	-	-	-	-	-
	No DIF	-	-	-	0	0	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	4	-	-	-	-	-	-
STEBvLTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	5	0	0	5	0	0	4	0	0	3	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0
EPvSPA	Favor Ref	-	-	-	0	0	1	-	-	-	-	-	-
	No DIF	-	-	-	0	0	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	4	-	-	-	-	-	-
EPvOTH	Favor Ref	-	-	-	0	0	0	-	-	-	-	-	-
	No DIF	-	-	-	0	1	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	3	1	-	-	-	-	-	-
OTHvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	3	0	0	3	0	0	2	0	0	1	0
	Favor Focal	0	1	1	0	2	0	0	2	0	0	2	0

Table 3.15.

Number of Low LC Items Changing DIF Significance or Direction After Conditioning on Linguistic Complexity – Mathematics Assessment

Analysis	Base Model DIF Direction	LEX			NP			RC			All Predictors		
		Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal
EPvEB	Favor Ref	-	-	-	0	0	5	-	-	-	-	-	-
	No DIF	-	-	-	0	1	1	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	1	-	-	-	-	-	-
EPvSTEB	Favor Ref	-	-	-	0	0	7	-	-	-	-	-	-
	No DIF	-	-	-	0	1	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	0	-	-	-	-	-	-
EPvLTEB	Favor Ref	-	-	-	0	0	4	-	-	-	-	-	-
	No DIF	-	-	-	0	1	2	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	1	-	-	-	-	-	-
STEBvLTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	5	0	0	8	0	0	34	0	0	11	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0
EPvSPA	Favor Ref	-	-	-	0	0	6	-	-	-	-	-	-
	No DIF	-	-	-	0	1	0	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	1	-	-	-	-	-	-
EPvOTH	Favor Ref	-	-	-	0	0	2	-	-	-	-	-	-
	No DIF	-	-	-	0	3	2	-	-	-	-	-	-
	Favor Focal	-	-	-	0	0	1	-	-	-	-	-	-
OTHvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	5	0	0	7	0	0	30	3	0	11	0
	Favor Focal	0	0	0	0	0	1	0	0	1	0	0	0

EPvEB

The omnibus test results for this analysis are presented in Table 3.16. The results reveal that adding lexical complexity factor scores or relative clause factor scores to the base model as a single LC predictor and their interactions with EB status does not significantly improve model fit, while complex noun phrase factor scores appear to improve model fit and influence item responses. Including relative clauses factor scores as a single LC predictor greatly worsens model fit. Due to only complex noun phrases improving model fit, models with multiple LC predictors were not created.

Table 3.16.

EPvEB Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-6044961	-	-	-	12090014	12090145
Base model	-6351874	-	-	-	12703930	12704189
LEX predictor	-6611590	4	-519432	N/A	13223370	13223640
NP predictor	-6170816	4	362116	< 0.001	12341822	12342092
RC predictor	-1773073000	4	-3533442252	N/A	3546146190	3546146460

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F1 presents the base model's item difficulties for the reference (EP) and focal (EB) groups and the differences in the item difficulties between groups (positive values indicate the item was more difficult for the focal group and negative values indicate the item was more difficult for the reference group). Table G1 presents the Rasch HGLM results for the base model and NP predictor model; the adjusted DIF estimates and confidence intervals are in Table G2. In the base model, 30 of 41 items exhibited significant DIF, 16 items had substantial DIF favoring EBs, 14 items had moderate DIF favoring EPs, and two items had substantial DIF favoring EPs.

All polytomous items favored EBs, but the thresholds for higher points (i.e., the thresholds for two, three, or four points) on these items favored EPs.

For the NP predictor model, 36 out of 41 items favored EBs when complex noun phrases were accounted for. All items that favored EBs or EPs in the base model favored EBs when complex noun phrases were accounted for. Six out of 11 items that exhibited non-significant DIF in the base model exhibited DIF favoring EBs when complex noun phrases were accounted for. Typically, items that exhibited DIF favoring EPs in the base model had low complex noun phrases factor scores, although many items favoring EBs in the base model also had low complex noun phrases factor scores. The items with the highest complex noun phrases factor scores favored EBs before and after accounting for complex noun phrases.

EPvSTEB

The omnibus test results for this analysis are presented in Table 3.17. The results reveal that adding either lexical complexity factor scores or relative clause factor scores to the base model and their interactions with STEB status does not significantly improve model fit, while complex noun phrase factor scores appear to improve model fit and influence item responses. Including relative clauses factor scores as a predictor greatly worsens model fit. Due to only complex noun phrases improving model fit, models with multiple LC predictors were not created.

Table 3.17.

EPvSTEB Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-6030024	-	-	-	12060140	12060270
Base model	-6247924	-	-	-	12496030	12496288
LEX predictor	-6515368	4	-534888	N/A	13030926	13031195
NP predictor	-6072379	4	351090	< 0.001	12144948	12145217
RC predictor	-1593316000	4	-3174136152	N/A	3186632190	3186632459

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F2 presents the base model's item difficulties for the reference (EP) and focal (STEB) groups and the differences in the item difficulties between groups. Table G3 presents the Rasch HGLM results for the base model and NP predictor model; the adjusted DIF estimates and confidence intervals are in Table G4. Generally, DIF direction and significance were the same for the EPvSTEB comparison group as it was for the EPvEB comparison group (meaning the same items changed DIF significance and direction after accounting for complex noun phrases for both comparison groups) with the exception of five items. These items appear to reflect differences in DIF detection in the base model with the exception of m38, an item with a high complex noun phrases factor score, exhibited moderate DIF favoring EPs in the base model for EPvEB and EPvSTEB, but after conditioning on complex noun phrases exhibited moderate DIF favoring EBs for EPvEB and non-significant DIF for EPvSTEB. Items m02 and m04 exhibited substantial DIF favoring EBs in the base model for EPvEB, but exhibited non-significant DIF in the base model for EPvSTEB. Items m07 and m41 exhibited non-significant DIF in the base model for EPvEB, but exhibited moderate DIF favoring EPs in the base model for EPvSTEB. The differences in DIF significance and direction were in the base model, but these differences were gone after accounting for complex noun phrases.

EPvLTEB

The omnibus test results for this analysis are presented in Table 3.18. The results reveal that adding lexical complexity factor scores or relative clause factor scores to the base model and their interactions with LTEB status does not significantly improve model fit, while complex noun phrase factor scores appear to improve model fit and influence item responses. Including relative clauses factor scores as a predictor greatly worsens model fit. Due to only complex noun phrases improving model fit, models with multiple LC predictors were not created.

Table 3.18.

EPvLTEB Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-6010309	-	-	-	12020710	12020840
Base model	-6127636	-	-	-	12255454	12255711
LEX predictor	-6406999	4	-558726	N/A	12814188	12814456
NP predictor	-5957879	4	339514	< 0.001	11915948	11916216
RC predictor	-1475565000	4	-2938874728	N/A	2951130190	2951130458

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F3 presents the base model's item difficulties for the reference (EB) and focal (LTEB) groups and the differences in the item difficulties between groups. Table G5 presents the Rasch HGLM results for the base model and NP predictor model; the adjusted DIF estimates and confidence intervals are in Table G6. Generally, DIF direction and significance were the same for the EPvLTEB comparison group as it was for the EPvEB comparison group with the exception of nine items. These items appear to reflect differences in DIF detection in the base model with the exception of two items. Item m22 exhibited non-significant DIF in the base model for EPvEB and EPvLTEB, but after conditioning complex noun phrases exhibited non-significant DIF for EPvEB and moderate DIF favoring LTEBs for EPvLTEB. Item m38, an item with a high complex noun phrases factor score, exhibited moderate DIF favoring EPs in the base model for EPvEB and EPvLTEB, but after conditioning on complex noun phrases exhibited moderate DIF favoring EBs for EPvEB and non-significant DIF for EPvLTEB. Items m03, m08, m17, m21, and m32 exhibited significant DIF favoring EPs in the base model for EPvEB, but non-significant DIF for EPvLTEB. Item m04 exhibiting substantial DIF favoring EBs in the base model for EPvEB, but non-significant DIF for EPvLTEB. The differences in DIF significance and direction were in the base model, but these differences were gone after accounting for complex noun phrases. Item m34 did not follow this pattern: for EPvEB this item exhibited non-significant DIF in the base model and the NP predictor model, but for EPvLTEB exhibited moderate DIF favoring EPs in the base model and moderate DIF favoring LTEBs in the NP predictor model.

STEBvLTEB

The omnibus test results for this analysis are presented in Table 3.19. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL) = 1{,}197.2$, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model ($-2(\Delta LL) = 18{,}881.6$, $\Delta df = 4$, $p < .001$). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how complex noun phrase factor scores influence item responses.

Table 3.19.

STEBvLTEB Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-333931.3	-	-	-	667954.6	668027.6
Base model	-335774.2	-	-	-	671730.4	671874.8
LEX predictor	-306466.4	4	58615.6	< 0.001	613122.8	613273.6
NP predictor	-324261.8	4	23024.8	< 0.001	648713.6	648864.4
RC predictor	-309840.2	4	51868.0	< 0.001	619870.4	620021.2
LEX + RC predictors	-305867.8	8	59812.8	< 0.001	611933.6	612090.7
All predictors	-296427.0	12	78694.4	< 0.001	593060.0	593223.5

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F4 presents the base model's item difficulties for the reference (STEB) and focal (LTEB) groups and the differences in the item difficulties between groups. Table G7 presents the Rasch HGLM results for the base model examining DIF between STEBs and LTEBs and the models including LC factor scores as item-level predictors of item responses; the adjusted DIF estimates and confidence intervals are in Table G8 and the covariance matrix for the all predictors model is in Table G15. The covariances between LC features and the intercept were high and positive with the exception of complex noun phrases; this LC factor had small negative covariance with the intercept, a small positive covariance with lexical complexity, and a negative moderate covariance with relative clauses. In the base model, no items exhibited significant DIF and there were no significant group differences in item estimates. Despite the significant interactions between LC predictors and LTEB status, that suggest lexical complexity and relative clauses, but not complex noun phrases, may influence item responses differently for STEBs and LTEBs, no items changed DIF direction or significance between the base model and any LC predictor models. However, accounting for lexical complexity or relative clauses factor scores leads to LTEBs having significantly higher ability estimates than STEBs; items with higher lexical complexity or relative clauses factor scores are easier for STEBs. Accounting for complex noun phrases factor scores did not lead to groups differences in ability estimates or effect of complex noun phrases on item responses. Accounting for all predictors led to no significant group differences in ability estimates or effects of any LC predictors on item responses, although lexical complexity significantly increases item difficulty and complex noun phrases significantly decreases item difficulty.

EPvSPA

The omnibus test results for this analysis are presented in Table 3.20. The results reveal that adding lexical complexity factor scores or relative clause factor scores to the base model and their interactions with SPA status does not significantly improve model fit, while complex noun phrase factor scores appear to improve model fit and influence item responses. Including relative clauses factor scores as a predictor greatly worsens model fit. Due to only complex noun phrases improving model fit, models with multiple LC predictors were not created.

Table 3.20.

EPvSPA Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-6000657	-	-	-	12001406	12001536
Base model	-6228968	-	-	-	12458118	12458376
LEX predictor	-6491925	4	-525914	N/A	12984040	12984309
NP predictor	-6050615	4	356706	< 0.001	12101420	12101689
RC predictor	-1667054000	4	-3321650064	N/A	3334108190	3334108459

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F5 presents the base model's item difficulties for the reference (EP) and focal (SPA) groups and the differences in the item difficulties between groups. Table G9 presents the Rasch HGLM results for the base model and NP predictor model; the adjusted DIF estimates and confidence intervals are in Table G10. Generally, DIF detection and DIF effect size were the same for the EPvSPA comparison group as it was for the EPvEB comparison group with the exception of five items; these items appear to reflect differences in DIF detection in the base model. Items m08 and m17 exhibited significant DIF favoring EPs in the base model for EPvEB, but non-significant DIF for EPvSPA. Items m25 and m45 exhibited non-significant DIF in the base model for EPvEB, but substantial DIF favoring EPs for EPvSPA. The differences in DIF significance and direction were in the base model, but these differences were gone after accounting for complex noun phrases. Item m34 did not follow this pattern: for EPvEB this item exhibited non-significant DIF in the base model and the NP predictor model, but for EPvSPA exhibited moderate DIF favoring EPs in the base model and moderate DIF favoring Spanish-speaking EBs in the NP predictor model.

EPvOTH

The omnibus test results for this analysis are presented in Table 3.21. The results reveal that adding lexical complexity factor scores or relative clause factor scores to the base model and their interactions with OTH status does not significantly improve model fit, while complex noun phrase factor scores appear to improve model fit and influence item responses. Including relative clauses factor scores as a predictor greatly worsens model fit. Due to only complex noun phrases improving model fit, models with multiple LC predictors were not created.

Table 3.21.

EPvOTH Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-6040114	-	-	-	12080320	12080450
Base model	-6152462	-	-	-	12305106	12305363
LEX predictor	-6432593	4	-560262	N/A	12865376	12865645
NP predictor	-5982913	4	339098	< 0.001	11966016	11966285
RC predictor	-1433833000	4	-2855361076	N/A	2867666190	2867666459

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F6 presents the base model's item difficulties for the reference (EP) and focal (OTH) groups and the differences in the item difficulties between groups. Table G11 presents the Rasch HGLM results for the base model and NP predictor model; the adjusted DIF estimates and confidence intervals are in Table G12. There were substantial differences in DIF direction and significance between the EPvEB and EPvOTH comparison groups, 14 out of 41 items had differences. The differences in items m02, m03, m04, and m06 appear to reflect differences in DIF detection in the base model. For EPvEB, items m02 and m04 exhibited substantial DIF favoring EBs and items m03 and m08 exhibited substantial DIF favoring EPs in the base model, but these items exhibited non-significant DIF in the base model for EPvOTH. For both comparison groups, items exhibited significant DIF favoring EPs after conditioning for complex noun phrases. The differences in items m01, m06, m21, m26, and m38 also reflect differences in DIF detection in the base model between comparison groups, but also differences in DIF significance after accounting for complex noun phrases. For EPvEB, items m01, m21, m26, and m38 exhibited moderate DIF favoring EPs and items m06 exhibited substantial DIF favoring EBs in the base model, but these items exhibited non-significant DIF in the base model for EPvOTH. For EPvEB, these items exhibited significant DIF favoring EBs after accounting for complex noun phrases, but for EPvOTH, these items exhibited non-significant DIF. While items m10, m33, m35, and m40 did not have differences in DIF direction or significance in the base model between comparison groups, in the NP predictor model, these items exhibit significant DIF favoring EBs for EPvEB but non-significant DIF for EPvOTH. Generally, less items exhibit significant DIF after accounting for complex noun phrases for EPvOTH compared to EPvEB.

OTHvSPA

The omnibus test results for this analysis are presented in Table 3.22. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL) = 12{,}087.2$, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model ($-2(\Delta LL) = 11{,}741.4$, $\Delta df = 4$, $p < .001$). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how complex noun phrase factor scores influence item responses.

Table 3.22.

OTHvSPA Omnibus Test Results – Mathematics Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-333818.6	-	-	-	667729.2	667802.2
Base model	-342514.7	-	-	-	685211.4	685355.8
LEX predictor	-309519.9	4	65989.6	< 0.001	619229.8	619380.6
NP predictor	-327927.4	4	29174.6	< 0.001	656044.8	656195.6
RC predictor	-310203.6	4	64622.2	< 0.001	620597.2	620748.0
LEX + RC predictors	-303476.3	8	78076.8	< 0.001	607150.6	607307.7
All predictors	-297605.6	12	89818.2	< 0.001	595417.2	595580.7

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F7 presents the base model's item difficulties for the reference (OTH) and focal (SPA) groups and the differences in the item difficulties between groups. Table G13 presents the Rasch HGLM results for the base model examining DIF between non-Spanish-speaking EBs and Spanish-speaking EBs and the models including LC factor scores as item-level predictors of item responses; the adjusted DIF estimates and confidence intervals are in Table G14 and the covariance matrix for the all predictors model is in Table G15. The covariances between LC features and the intercept were high and positive with the exception of complex noun phrases; this LC factor had small negative covariance with the intercept, a small positive covariance with lexical complexity, and a negative moderate covariance with relative clauses. In the base model, non-Spanish-speaking EBs had significantly higher ability estimates than Spanish-speaking EBs, with three items exhibiting significant DIF: m09, m14 and m35 exhibited moderate DIF favoring Spanish-speaking EBs.

There are significant interactions between LC predictors and SPA status, which suggests lexical complexity and relative clauses, but not complex noun phrases, may influence item responses differently for non-Spanish-speaking EBs and Spanish-speaking EBs. Accounting for lexical complexity factor scores leads to Spanish-speaking EBs having significantly higher ability estimates than non-Spanish-speaking EBs, while accounting relative clauses factor scores leads to non-Spanish-speaking EBs having significantly higher ability estimates than Spanish-speaking EBs. However, items with higher lexical complexity or relative clauses factor scores are easier for non-Spanish-speaking EBs. Accounting for complex noun phrases factor scores does not lead to groups differences in ability estimates, but items with higher complex noun phrases factor scores are easier for non-Spanish-speaking EBs. Accounting for all predictors leads to no significant group differences in ability estimates or effects of any LC predictors on item responses, although lexical complexity and relative clauses significantly increase item difficulty and complex noun phrases significantly decreases item difficulty.

Generally, items did not change DIF significance or direction between the base model and any LC predictor model. In the base model, m09 exhibited moderate DIF favoring Spanish-speaking EBs; after accounting for LC factor score predictors, this item exhibited substantial DIF favoring Spanish-speaking EBs in the LEX predictor, NP predictors, RC predictor, and all predictors models. Items m14 and m35 exhibited moderate DIF favoring Spanish-speaking EBs; these DIF estimates were non-significant for all four LC predictor models. Items m02, m16, and m33 were also exceptions; these items went from exhibiting non-significant DIF in the base model to exhibiting moderate DIF favoring Spanish-speaking EBs in the RC predictor models.

MCAS Biology Results

Summary of Subgroup Results for the Biology Assessment

Due to conducting many HGLMs, summaries of the results are presented and discussed first, starting with model fit comparisons in Table 3.23. Significant changes in log likelihood and lower AIC and BIC values indicated improvements in model fit. For all models, when changes in log likelihood was significant, AIC and BIC were lower, and when changes in log likelihood were significant, AIC and BIC were higher, therefore model fit results agreed with each other. The results of the omnibus tests and AIC and BIC comparisons demonstrate the base model assessing DIF did not improve model fit compared to the comparison model that did not assess DIF. As with the mathematics assessment, analyses continued as identifying DIF ensures valid interpretations of scores for all test-takers.

For all comparison groups, each LC predictor model improved model fit compared to the base model. As the LEX predictor and RC predictor models improved fit the most (for all comparison groups), a multiple LC predictor model with both of these factors was examined next, this model improved fit for all comparison groups. A model with all three LC predictors improved model fit compared to the model with the LEX and RC predictors, suggesting that the inclusion of these LC factor scores explains item responses on this biology assessment. Interestingly, the LEX and RC predictors models and the all predictors models for STEBvLTEB, EPvSPA, and OTHvSPA improved model fit compared to the comparison model, and the LEX predictor model for STEBvLTEB improved model fit compared to the comparison model. As with the mathematics assessment, the inclusion of these LC factor scores may explain EBs' item responses in particular. Specifics of the omnibus tests and changes in AIC and BIC for the inclusion of each LC factor and are presented in the results for each comparison group.

Table 3.23.

Summary of Model Fit Improvement for each Comparison Group – Biology Assessment

Comparison Group	Improved Model Fit Compared to:	Base	LEX	NP	RC	LEX + RC	All predictors
EPvEB	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
EPvSTEB	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
EPvLTEB	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
STEBvLTEB	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
EPvSPA	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
EPvOTH	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓
OTHvSPA	Comparison	X	-	-	-	-	-
	Base	-	✓	✓	✓	-	-
	Any Single LC Predictor	-	-	-	-	✓	-
	LEX + RC	-	-	-	-	-	✓

Note: "✓" indicates significant changes in -2 log likelihood and lower AIC and BIC values that led to a judgement of improved model fit. "X" indicates non-significance.

Table 3.24 presents the significance of the intercept's interaction with focal group status (γ_{01}) for each comparison group for the biology assessment. Positive interactions indicated the focal group had higher ability estimates; negative interactions indicated the reference group had higher ability estimates. For the EP versus EB comparison groups, EPs consistently had higher ability estimates in the base model, but when a single LC factor score predictor and its interaction with focal group status were included in the model, EBs had higher ability estimates. The exception to this was the NP predictor model for EPvLTEB; there were no significant group differences in ability between EPs and LTEBs. For the all predictors models, EBs had significantly higher ability estimates than EPs for all EP versus EB comparison groups.

For the EB versus EB comparison groups, there were no significant group differences in ability estimates for STEBvLTEB in the base model, but non-Spanish-speakers had higher ability estimates than Spanish-speakers in the OTHvSPA base model. When lexical complexity or complex noun phrases and their interaction with focal group status were accounted for in the EB versus EB comparison group models, there were no significant group differences in ability estimates. However, when relative clauses and its interaction with focal group status were accounted for in the EB versus EB comparison group models, there were no significant group differences in ability estimates for STEBvLTEB, but non-Spanish-speaking EBs had higher ability estimates than Spanish-speaking EBs for OTHvSPA. For the all predictors models, there were no significant group differences in ability estimates.

Table 3.24.

Significance of the Intercept's Interaction with Focal Group Status (γ_{01}) for each Comparison Group – Biology Assessment

Comparison Group	**Base Model**	**LEX**	**NP**	**RC**	**All Predictors**
EPvEB	Favors EPs ***	Favors EBs ***	Favors EBs ***	Favors EPs ***	Favors EBs ***
EPvSTEB	Favors EPs ***	Favors STEBs ***	Favors STEBs ***	Favors EPs ***	Favors STEBs **
EPvLTEB	Favors EPs ***	Favors LTEBs ***	0.121	Favors EPs ***	Favors LTEBs *
STEBvLTEB	0.368	0.948	0.947	0.501	0.697
EPvSPA	Favors EPs ***	Favors SPAs ***	Favors SPAs ***	Favors EPs ***	Favors SPAs **
EPvOTH	Favors EPs ***	Favors OTHs ***	Favors OTHs ***	Favors EPs ***	Favors OTHs *
OTHvSPA	Favors OTHs **	0.293	0.684	Favors OTHs *	0.941

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$. If γ_{01} was not significant, *p*-values were listed instead.

Table 3.25 presents the significance of LC Factor predictors' main effects (γ_{s0}) and their interactions with focal group status (γ_{s1}) for each comparison group for the biology assessment. Negative interactions indicated items with higher LC factor scores were easier for the focal group; positive interactions indicated items with higher LC factor scores were easier for the reference group. For all comparison groups, the significant positive main effect of lexical complexity indicated that items with higher lexical complexity factor scores are associated with increased ability estimates than items with lower lexical complexity factor scores. For all comparison groups, the significant positive main effect of complex noun phrases indicated that items with higher complex noun phrases factor scores are associated with increased ability estimates than items with lower complex noun phrases factor scores. For all comparison groups, the significant negative main effect of relative clauses indicated that items with higher relative clauses factor scores are associated with decreased ability estimates than items with lower relative clauses factor scores.

For the EP versus EB comparison groups, items with higher lexical complexity or complex noun phrases factor scores were consistently easier for EPs than for EBs with the exception of the NP predictor model for EPvLTEB. Items with higher relative clauses factor scores were consistently easier for EBs than for EPs. However, differences between the reference and focal groups emerged when examining the EB versus EB comparison groups. For all LC predictor models, there were no group differences in how LC factor scores influenced item responses for either EB versus EB comparison group.

Table 3.25.

Significance of LC Factor Predictors (γ_{s0}) and Interactions with Focal Group Status (γ_{s1}) for each Comparison Group for Single LC Predictor Models – Biology Assessment

Comparison Group	LEX		NP		RC	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
EPvEB	***	Favors EPs ***	***	Favors EPs ***	***	Favors EBs ***
EPvSTEB	***	Favors EPs ***	***	Favors EPs ***	***	Favors STEBs ***
EPvLTEB	***	Favors EPs ***	***	0.065	***	Favors LTEBs ***
STEBvLTEB	***	0.909	***	0.974	***	0.574
EPvSPA	***	Favors EPs ***	***	Favors EPs ***	***	Favors SPAs ***
EPvOTH	***	Favors EPs ***	***	Favors EPs ***	***	Favors OTHs **
OTHvSPA	***	0.222	**	0.773	***	0.065

Note: *** = $p < .001$, ** = $p < .01$. If γ_{s1} was not significant, *p*-values were listed instead.

Table 3.26 presents the significance of LC factor predictors' main effects and their interactions with focal group status for all comparison groups for the biology assessment when all LC predictors are included in the model. For all EP versus EB comparison groups, the significant positive main effect of lexical complexity indicates that items with higher lexical complexity factor scores are associated with increased ability estimates than items with lower lexical complexity factor scores, the positive main effect of complex noun phrases indicates that items with higher complex noun phrases factor scores are associated with increased ability estimates than items with lower complex noun phrases factor scores, and the significant negative main effect of relative clauses indicates that items with higher relative clauses factor scores are associated with decreased ability estimates than items with lower relative clauses factor scores.

For both EB versus EB comparison groups, the non-significant main effects of lexical complexity and complex noun phrases indicates that lexical complexity and complex noun phrases did not influence ability estimates, and the significant negative main effect of relative clauses indicates that items with higher relative clauses factor scores are associated with decreased ability estimates than items with lower relative clauses factor scores.

Items with higher lexical complexity factor scores were consistently easier for EPs than for EBs. There were no significant group differences in how complex noun phrases factor scores influenced item responses for any EP versus EB comparison group. However, there were differences in the significance of the relative clauses factor scores interaction with focal group status. The interaction was significant for EPvEB, EPvSTEB, and EPvSPA; items with higher relative clauses factor scores were consistently easier for EBs, STEBs, and Spanish-speaking EBs than for EPs. The interaction was not significant for EPvLTEB and EPvOTH; there were no group differences in how relative clauses factor scores influenced item responses for these comparison groups. There were no significant interactions between any LC predictor and focal group status for either EB versus EB comparison group.

Table 3.26.

Significance of LC Factor Predictors and Interactions with Focal Group Status for EP Versus EB Comparison Groups for All Predictor Models – Biology Assessment

Comparison Group	LEX		NP		RC	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
EPvEB	***	Favors EPs ***	***	0.162	***	Favors EBs *
EPvSTEB	***	Favors EPs ***	***	0.265	***	Favors STEBs *
EPvLTEB	***	Favors EPs *	***	0.310	***	0.349
STEBvLTEB	0.164	0.953	0.137	0.521	**	0.910
EPvSPA	***	Favors EPs ***	***	0.437	***	Favors SPAs *
EPvOTH	***	Favors EPs ***	***	0.169	***	0.215
OTHvSPA	0.081	0.247	0.783	0.420	*	0.550

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$. If γ_{s1} was not significant, *p*-values were listed instead.

The number of items changing DIF significance or direction with the inclusion of an LC factor and its interaction with focal group status for biology assessment are presented in Table 3.27. This section provides an overview of the items changing DIF significance or direction with specific results presented for each comparison group below, which includes the differences between the EPvEB models and EP versus subgroups of EB models.

For the EPvEB comparison groups, many items that exhibited DIF favoring EBs in the base model were the polytomous items on the biology assessment. Examination of the thresholds revealed meeting these higher-point thresholds exhibited DIF in favor of EPs or were non-significant. Therefore, these items that favored EBs in the base model were indicators that EBs had an easier time achieving the one-point threshold than EPs. When accounting for LC factor scores, these items generally did not change DIF significance or direction. The exceptions were for the RC predictor model, where the polytomous items exhibited substantial DIF favoring EPs after accounting for relative clauses (except for one item that exhibited non-significant DIF for EPvOTH), and the NP predictor model for EPvLTEB, where some polytomous items exhibited non-significant DIF after accounting for complex noun phrases.

For the dichotomous items, accounting for lexical complexity lead to many items favoring EBs and EB subgroups, although accounting for complex noun phrases lead to virtually all items exhibiting non-significant DIF. Accounting for relative clauses led to mixed results, with most items exhibiting non-significant DIF, although many items that favored EBs in the base model changed direction to favoring EPs in the RC predictor model. Accounting for all LC predictors also led to mixed results, with items exhibiting non-significant DIF or significant DIF favoring EBs. Generally, the EPvOTH comparison group had less items in the base model exhibiting significant DIF.

Tables 3.28 and 3.29 show the items changing DIF significance or direction with the inclusion of an LC factor and its interaction with focal group status for items with high or low LC factor scores, respectively. For the all predictors models, only items with two or more high LC factors and no low LC factors (Table 3.28) or two or more low LC factors and no high LC factors (Table 3.29) were considered. Items were considered as having a high LC factor score when the LC factor score was greater than one standard deviation above the mean. Items were considered as having a low LC factor score when the LC factor score was greater than one standard deviation below the mean for lexical complexity or was the lowest LC factor score value for complex noun phrases and relative clauses.

For the EP versus EB comparison groups, the items with high lexical complexity factors scores, b01, b09, b10, b12, b17, b27, b44, exhibited substantial DIF favoring EBs after lexical complexity was accounted for. The items with low lexical complexity, b11, b18, b19, b26, b29, and b39, exhibited substantial DIF favoring EBs after accounting for lexical complexity. The items with high complex noun phrases factor scores, b17, b21, b24, b28, b36, b40, and b42, exhibited non-significant DIF after complex noun phrases was accounted for. The items with low complex noun phrases, b15, b19, b22, b26, b33, and b39 exhibited non-significant DIF after complex noun phrases was accounted for. The items with high relative clauses, b01, b17, b28, and b36, exhibited non-significant DIF after relative clauses were accounted for. The items with low relative clauses, b03, b04, b05, b06, b07, b11, b13, b15, b16, b18, b19, b20, b21, b22, b25, b26, b29, b30, b31, b32, b33, b35, b39, and b42, generally exhibited non-significant DIF favoring EBs after accounting for relative clauses, with the exception of some items that favored EBs in the model that exhibited DIF favoring EPs after accounting for relative clauses. The EPvOTH comparison group has more low relative clauses items exhibited non-significant DIF, but this comparison group has less items overall with significant DIF than other comparison groups.

The items with high factor scores (at least two high LC factor scores and no low LC factor scores) in the all predictors models, b01, b17, b28, and b36 exhibited non-significant DIF after accounting for LC features. The items with low factor scores (at least two low LC factor scores and no high LC factor scores) in the all predictors models, b11, b15, b18, b19, b22, b26, b29, b33, and b39, had different patterns of DIF significance between comparison groups. For EPvEB, EPvSTEB, EPvLTEB, and EPvOTH, the items were split between exhibited significant DIF favoring EBs or non-significant DIF, but for EPvSPA, the items exhibited significant DIF favoring EBs. To summarize the EP versus EB comparison group results, high or low factor scores did not influence changes in DIF significance or direction for the LEX predictor or NP predictor models, while high factor scores led to non-significant DIF in the RC predictor and all predictors models and low factor scores led to mixed results in DIF in the RC predictor and all predictors models.

For the EB versus EB comparison groups, items generally did not exhibit significant DIF across all models. For STEBvLTEB, no items exhibited significant DIF in the base model and the inclusion of any LC predictor did not change items' DIF direction or significance. For OTHvSPA, the few items that did exhibit DIF in the base model favored Spanish-speaking EBs and accounting for any LC predictor generally led to these items exhibiting non-significant DIF, although one item changed direction to favoring non-Spanish-speaking EBs after accounting for relative clauses. Interestingly, the few items that exhibited DIF in the base model tended to have high LC factor scores.

Table 3.27.

Number of Items Changing DIF Significance or Direction After Conditioning on Linguistic Complexity – Biology Assessment

Analysis	Base Model DIF Direction	LEX			NP			RC			All Predictors		
		Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal
EPvEB	Favor Ref	0	0	5	0	5	0	0	5	0	0	0	5
	No DIF	0	1	23	0	24	0	0	24	0	0	9	15
	Favor Focal	0	0	16	0	11	5	8	8	0	0	4	12
EPvSTEB	Favor Ref	0	0	3	0	3	0	0	3	0	0	0	3
	No DIF	0	1	25	0	26	0	0	26	0	0	12	15
	Favor Focal	0	0	16	0	10	6	8	8	0	0	5	10
EPvLTEB	Favor Ref	0	0	6	0	6	0	0	6	0	0	2	4
	No DIF	0	1	27	0	28	0	0	28	0	0	11	17
	Favor Focal	0	0	11	0	8	3	7	4	0	0	4	7
STEBv LTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	45	0	0	45	0	0	45	0	0	45	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0
EPvSPA	Favor Ref	0	0	6	0	6	0	0	6	0	0	0	6
	No DIF	0	1	24	0	25	0	0	25	0	0	5	20
	Favor Focal	0	0	14	0	7	7	8	6	0	0	5	9
EPvOTH	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	1	31	0	32	0	0	32	0	0	22	10
	Favor Focal	0	0	13	0	8	5	5	8	0	0	5	8
OTHvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	42	0	0	42	0	0	42	0	0	42	0
	Favor Focal	0	3	0	0	3	0	1	2	0	0	3	0

Table 3.28.

Number of High LC Items Changing DIF Significance or Direction After Conditioning on Linguistic Complexity – Biology Assessment

Analysis	Base Model DIF Direction	LEX			NP			RC			All Predictors		
		Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal
EPvEB	Favor Ref	0	0	1	0	0	0	0	0	0	0	0	0
	No DIF	0	0	2	0	3	0	0	1	0	0	1	0
	Favor Focal	0	0	4	0	4	0	0	3	0	0	3	0
EPvSTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	0	3	0	3	0	0	1	0	0	1	0
	Favor Focal	0	0	4	0	4	0	0	3	0	0	3	0
EPvLTEB	Favor Ref	0	0	1	0	0	0	0	0	0	0	0	0
	No DIF	0	0	4	0	4	0	0	2	0	0	2	0
	Favor Focal	0	0	2	0	3	0	0	2	0	0	2	0
STEBvLTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	7	0	0	7	0	0	4	0	0	4	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0
EPvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	0	4	0	3	0	0	1	0	0	1	0
	Favor Focal	0	0	3	0	4	0	0	3	0	0	3	0
EPvOTH	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	0	3	0	4	0	0	1	0	0	1	0
	Favor Focal	0	0	4	0	3	0	0	3	0	0	3	0
OTHvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	7	0	0	7	0	0	4	0	0	4	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.29.

Number of Low LC Items Changing DIF Significance or Direction After Conditioning on Linguistic Complexity – Biology Assessment

Analysis	Base Model DIF Direction	LEX			NP			RC			All Predictors		
		Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal	Favor Ref	No DIF	Favor Focal
EPvEB	Favor Ref	0	0	1	0	1	0	0	4	0	0	0	1
	No DIF	0	1	4	0	6	0	0	15	0	0	4	4
	Favor Focal	0	0	1	0	0	0	5	1	0	0	0	1
EPvSTEB	Favor Ref	0	0	0	0	0	0	0	3	0	0	0	0
	No DIF	0	1	5	0	7	0	0	16	0	0	4	5
	Favor Focal	0	0	1	0	0	0	5	1	0	0	0	1
EPvLTEB	Favor Ref	0	0	0	0	0	0	0	4	0	0	0	0
	No DIF	0	1	5	0	7	0	0	16	0	0	3	6
	Favor Focal	0	0	1	0	0	0	4	1	0	0	0	1
STEBvLTEB	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	7	0	0	7	0	0	25	0	0	10	0
	Favor Focal	0	0	0	0	0	0	0	0	0	0	0	0
EPvSPA	Favor Ref	0	0	1	0	1	0	0	5	0	0	0	1
	No DIF	0	1	4	0	6	0	0	14	0	0	1	7
	Favor Focal	0	0	1	0	0	0	5	1	0	0	0	1
EPvOTH	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	1	5	0	7	0	0	20	0	0	6	3
	Favor Focal	0	0	1	0	0	0	2	3	0	0	0	1
OTHvSPA	Favor Ref	0	0	0	0	0	0	0	0	0	0	0	0
	No DIF	0	7	0	0	7	0	0	24	0	0	10	0
	Favor Focal	0	0	0	0	0	0	0	1	0	0	0	0

EPvEB

The omnibus test results for this analysis are presented in Table 3.30. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL) = 31{,}662$, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model ($-2(\Delta LL) = 5{,}670$, $\Delta df = 4$, $p < .001$). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how LC factor scores influence item responses.

Table 3.30.

EPvEB Omnibus Test Results – Biology Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-1531863	-	-	-	3063824	3063933
Base model	-1709705	-	-	-	3419604	3419821
LEX predictor	-1693812	4	31786	< 0.001	3387826	3388052
NP predictor	-1707010	4	5390	< 0.001	3414222	3414448
RC predictor	-1702181	4	15048	< 0.001	3404564	3404790
LEX + RC predictors	-1677981	8	63448	< 0.001	3356172	3356407
All predictors	-1675146	12	69118	< 0.001	3350510	3350753

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F8 presents the base model's item difficulties for the reference (EP) and focal (EB) groups and the differences in the item difficulties between groups (positive values indicate the item was more difficult for the focal group and negative values indicate the item was more difficult for the reference group). Table G16 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G17 and the covariance matrix for the all predictors model is in Table G30. The covariances between LC features and the intercept were large and positive with the exception of relative clauses; this LC factor had large negative covariances with the intercept, lexical complexity, and complex noun phrases. All LC predictors in the single LC predictor models had significant interactions with EB status ($p < .05$).

In the base model, 21 of 44 items tested for DIF exhibited significant substantial DIF favoring EBs, 16 items had substantial DIF favoring EBs and 5 items had moderate DIF favoring EPs. All polytomous items favored EBs, but the thresholds for higher points (i.e., the thresholds for two, three, or four points) on these items favored EPs. These items continued to favor EBs for the LEX predictor, NP predictor, and all predictors models, but not the RC predictor model; these items exhibited substantial DIF favoring EPs after accounting for relative clauses.

For the LEX predictor model, 44 out of 44 items favored EBs when lexical complexity was accounted for regardless of DIF direction or significance in the base model. For the NP predictor model, all dichotomous items exhibited non-significant DIF when complex noun phrases were accounted for, regardless of DIF direction or significance in the base model; the polytomous items remained favoring EBs. For the RC predictor model, most dichotomous items exhibited non-significant DIF when relative clauses were accounted for, with three dichotomous items and all five polytomous items that favored EBs in the base model favoring EPs in the RC predictor model. When accounting for all LC predictors, results were more mixed. In the all predictors model, lexical complexity and relative clauses factor scores had significant interactions with EB status ($p < .05$), but complex noun phrases factor scores did not ($p = .162$). While nine items that exhibited non-significant DIF in the base model continued to exhibit non-significant DIF after accounting for all LC predictors, fifteen items that exhibited non-significant DIF in the base model exhibited significant DIF favoring EBs after accounting for all LC predictors. Of the sixteen items favoring EBs in the base model, four items exhibited non-significant DIF and twelve items exhibited significant DIF favoring EBs when accounting for all LC predictors. All five items that favored EPs in the base models favored EBs when accounting for all LC predictors.

EPvSTEB

The omnibus test results for this analysis are presented in Table 3.31. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model (-2(ΔLL) = 30,836, Δdf = 4, p < .001). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model (-2(ΔLL) = 5,622, Δdf = 4, p < .001). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how LC factor scores influence item responses.

Table 3.31.

EPvSTEB Omnibus Test Results – Biology Assessment

Model	**"LL"**	**Δdf**	**-2(ΔLL)**	***p*-value**	**AIC**	**BIC**
Comparison model	-1525078	-	-	-	3050254	3050363
Base model	-1667054	-	-	-	3334302	3334517
LEX predictor	-1651741	4	30626	< 0.001	3303684	3303908
NP predictor	-1664404	4	5300	< 0.001	3329010	3329234
RC predictor	-1659640	4	14828	< 0.001	3319482	3319706
LEX + RC predictors	-1636323	8	61462	< 0.001	3272856	3273089
All predictors	-1633512	12	67084	< 0.001	3267242	3267484

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F9 presents the base model's item difficulties for the reference (EP) and focal (STEB) groups and the differences in the item difficulties between groups. Table G18 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G19 and the covariance matrix for the all predictors model is in Table G30. The covariances between LC features and the intercept were large and positive with the exception of relative clauses; this LC factor had large negative covariances with the intercept, lexical complexity, and complex noun phrases. Generally, DIF detection and DIF effect size were the same for the EPvSTEB comparison group as it was for the EPvEB comparison group with the exception of five items. Items b10 and b26 exhibited moderate DIF favoring EPs in the base model for EPvEB, but non-significant DIF in the base model for EPvSTEB. Item b10 also had different DIF estimates between comparison groups for the all predictors model: after accounting for all LC predictors, b10 exhibited moderate DIF favoring EBs for EPvEB and non-significant DIF for EPvSTEB. Three items had the same direction of DIF in the base model for both comparison groups, but differences in DIF direction after accounting for LC predictors: b19 and b20 exhibited moderate DIF favoring EBs in the all predictors model for EPvEB, but no significant DIF for EPvSTEB, and b25 exhibited moderate DIF favoring STEBs in the NP predictor model for EPvSTEB, but no significant DIF for EPvEB.

EPvLTEB

The omnibus test results for this analysis are presented in Table 3.32. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL)$ = 30,836, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was not added to the "LEX + RC predictors" model as complex noun phrases factor scores in the NP predictor model did not have a significant interaction with focal group status (discussed later in this section). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how LC factor scores influence item responses.

Table 3.32.

EPvLTEB Omnibus Test Results – Biology Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-1528967	-	-	-	3058032	3058140
Base model	-1594832	-	-	-	3189858	3190071
LEX predictor	-1581781	4	170546	< 0.001	3163764	3163986
NP predictor	-1592546	4	149016	< 0.001	3185294	3185516
RC predictor	-1587688	4	158732	< 0.001	3175578	3175800
LEX + RC predictors	-1567358	8	199392	< 0.001	3134926	3135157
All predictors	-1564641	12	204826	< 0.001	3129500	3129739

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F10 presents the base model's item difficulties for the reference (EP) and focal (LTEB) groups and the differences in the item difficulties between groups. Table G20 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G21 and the covariance matrix for the all predictors model is in Table G30. The covariances between LC features and the intercept were large and positive with the exception of relative clauses; this LC factor had large negative covariances with the intercept, lexical complexity, and complex noun phrases. There were many differences in DIF direction and significance between the EPvLTEB and EPvEB comparison groups. Twenty-four of 41 items had changes in DIF direction or significance, 14 of these 24 items reflected differences in which items were flagged for significant DIF in the base model. Items b01, b08, b09, b13, b34, and b40 exhibited substantial DIF favoring EBs and b05, b06, and b26 exhibited moderate DIF favoring EPs in the base model for EPvEB, but exhibited non-significant DIF in the base model for EPvLTEB. Items b04, b07, b35, and b43 exhibited moderate DIF favoring EPs and b37 exhibited substantial DIF favoring LTEBs in the base model for EPvLTEB, but exhibited non-significant DIF in the base model for EPvEB.

Ten items had the same direction of DIF in the base model for both comparison groups, but differences in DIF direction after accounting for LC predictors. For the NP predictor model, three items (b12, b32, and b44) that favored EBs in the base model remained favoring EBs after accounting for complex noun phrases for EPvEB, but these items exhibited non-significant DIF for EPvLTEB; one item (b11) that favored EBs in the base model remained favoring EBs remained favoring EBs after accounting for complex noun phrases for EPvLTEB, but this item exhibited non-significant DIF for EPvEB. For the RC predictor model, one item (b25), exhibited substantial DIF favoring EBs in the base model, but after accounting for relative clauses exhibited moderate DIF favoring EPs for EPvEB and non-significant DIF for EPvLTEB. For the all predictors model, six items (b10, b14, b27, b30, b39, and b44) differed in whether they exhibited moderate DIF favoring EBs or non-significant DIF after accounting for all LC predictors, but no clear pattern emerged. For the LEX predictor model, there were no changes in DIF direction between comparison groups after accounting for lexical complexity.

STEBvLTEB

The omnibus test results for this analysis are presented in Table 3.33. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model (-2(ΔLL) = 4,261.2, Δdf = 4, p < .001). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model (-2(ΔLL) = 74.2, Δdf = 4, p < .001). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how complex noun phrase factor scores influence item responses.

Table 3.33.

STEBvLTEB Omnibus Test Results – Biology Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-162059.7	-	-	-	324217.4	324280.3
Base model	-163645.2	-	-	-	327484.4	327608.9
LEX predictor	-159814.6	4	7661.2	< 0.001	319831.2	319960.9
NP predictor	-163355.1	4	580.2	< 0.001	326912.2	327041.9
RC predictor	-163162.6	4	965.2	< 0.001	326527.2	326656.9
LEX + RC predictors	-157684.0	8	11922.4	< 0.001	315578.0	315712.8
All predictors	-157646.9	12	11996.6	< 0.001	315511.8	315651.7

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F11 presents the base model's item difficulties for the reference (STEB) and focal (LTEB) groups and the differences in the item difficulties between groups. Table G22 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G23 and the covariance matrix for the all predictors model is in Table G30. The covariance between lexical complexity and the intercept is large and positive, between complex noun phrases and the intercept is moderate and positive, between complex noun phrases and lexical complexity is small and negative, between relative clauses and the intercept is large and negative, between relative clauses and lexical complexity is large and negative, and between relative clauses and relative clauses is moderate and negative. Across all models, no items exhibited significant DIF and there were no significance group differences in abilities estimates. In addition, there were no significant interactions between any LC predictors and LTEB status.

EPvSPA

The omnibus test results for this analysis are presented in Table 3.34. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL) = 30{,}758$, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model ($-2(\Delta LL) = 5{,}634$, $\Delta df = 4$, $p < .001$). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how LC factor scores influence item responses.

Table 3.34.

EPvSPA Omnibus Test Results – Biology Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-1509327	-	-	-	3018752	3018861
Base model	-1655817	-	-	-	3311828	3312043
LEX predictor	-1641224	4	29186	< 0.001	3282650	3282874
NP predictor	-1653300	4	5034	< 0.001	3306802	3307026
RC predictor	-1648402	4	14830	< 0.001	3297006	3297230
LEX + RC predictors	-1625845	8	59944	< 0.001	3251900	3252133
All predictors	-1623028	12	65578	< 0.001	3246274	3246515

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F12 presents the base model's item difficulties for the reference (EP) and focal (SPA) groups and the differences in the item difficulties between groups. Table G24 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G25 and the covariance matrix for the all predictors model is in Table G30. The covariances between LC features and the intercept were large and positive with the exception of relative clauses; this LC factor had large negative covariances with the intercept, lexical complexity, and complex noun phrases. There were many differences in DIF direction and significance between the EPvSPA and EPvEB comparison groups. Fourteen of 41 items had changes in DIF direction or significance, five of these 14 items reflected differences in which items were flagged for significant DIF in the base model. Items b09 and b34 exhibited substantial DIF favoring EBs and b10 exhibited moderate DIF favoring EPs in the base model for EPvEB, but exhibited non-significant DIF in the base model for EPvSPA. Items b35 and b41 exhibited moderate DIF favoring EPs in the base model for EPvSPA, but exhibited non-significant DIF in the base model for EPvEB.

Nine items had the same direction of DIF in the base model for both comparison groups, but differences in DIF direction after accounting for LC predictors. For the NP predictor model, two items (b11 and b25) that favored EBs in the base model remained favoring EBs after accounting for complex noun phrases for EPvSPA, but these items exhibited non-significant DIF for EpvEP. For the all predictors model, six items (b07, b22, b30, b33, b39, and b43) exhibited non-significant DIF after accounting for all LC predictors for EPvEB, but exhibited moderate DIF favoring Spanish-speaking EBs for EPvSPA. One item (b40) exhibited moderate DIF favoring EBs after accounting for all LC predictors for EPvEB, but exhibited non-significant DIF for EPvSPA. For the LEX predictor and RC predictor models, there were no changes in DIF direction between comparison groups after accounting for lexical complexity or relative clauses.

EPvOTH

The omnibus test results for this analysis are presented in Table 3.35. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL)$ = 29,508, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model ($-2(\Delta LL)$ = 5,6510, $\Delta df = 4$, $p < .001$). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how LC factor scores influence item responses.

Table 3.35.

EPvOTH Omnibus Test Results – Biology Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-1543037	-	-	-	3086172	3086280
Base model	-1608499	-	-	-	3217192	3217406
LEX predictor	-1595007	4	26984	< 0.001	3190216	3190438
NP predictor	-1606062	4	4874	< 0.001	3212326	3212548
RC predictor	-1601342	4	14314	< 0.001	3202886	3203108
LEX + RC predictors	-1580253	8	56492	< 0.001	3160716	3160947
All predictors	-1577498	12	62002	< 0.001	3155214	3155454

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F13 presents the base model's item difficulties for the reference (EP) and focal (OTH) groups and the differences in the item difficulties between groups. Table G26 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G27 and the covariance matrix for the all predictors model is in Table G30. The covariances between LC features and the intercept were large and positive with the exception of relative clauses; this LC factor had large negative covariances with the intercept, lexical complexity, and complex noun phrases. There were many differences in DIF direction and significance between the EPvOTH and EPvEB comparison groups. Nineteen of 41 items had changes in DIF direction or significance, eight of these 19 items reflected differences in which items were flagged for significant DIF in the base model. Items b08, b13, and b40 exhibited substantial DIF favoring EBs and items b03, b05, b06, b10, and b26 exhibited moderate DIF favorings EPs in the base model for EPvEB, but exhibited non-significant DIF in the base model for EPvOTH.

Eleven items had the same direction of DIF in the base model for both comparison groups, but differences in DIF direction after accounting for LC predictors. For the RC predictor model, three items (b11, b12, and b42) that favored EBs in the base model switched to favoring EPs after accounting for relative clauses for EPvEB, but these items exhibited non-significant DIF for EPvOTH. For the all predictors model, eight items (b09, b14, b19, b20, b31, b35, and b41) exhibited significant DIF favoring EBs after accounting for all LC predictors for EPvEB, but exhibited non-significant DIF for EPvOTH. For the LEX predictor and NP predictor models, there were no changes in DIF direction between comparison groups after accounting for lexical complexity or complex noun phrases.

OTHvSPA

The omnibus test results for this analysis are presented in Table 3.36. The results reveal that adding any of the three linguistic feature factors scores significantly improves model fit. As the LEX predictor model was the best fitting of the single LC predictor models, the LC factor from the next best-fitting model, the RC predictor model, was added to the LEX predictor model to determine if the inclusion of an additional LC predictor improved model fit ("LEX + RC predictors" model). This model fit significantly better than the LEX predictor model ($-2(\Delta LL)$ = 4,206.2, $\Delta df = 4$, $p < .001$). The last LC predictor, complex noun phrases, was added to the "LEX + RC predictors" model to determine if the inclusion of all LC predictors ("All predictors" model) significantly improved model fit. This model fit significantly better than the "LEX + RC predictors" model ($-2(\Delta LL) = 69.6$, $\Delta df = 4$, $p < .001$). AIC and BIC were lowest for the all predictors model. These results suggest that linguistic complexity, complex noun phrases, and relative clauses factor scores do influence item responses, but looking at the specific model results will reveal if there are group differences in how complex noun phrase factor scores influence item responses.

Table 3.36.

OTHvSPA Omnibus Test Results – Biology Assessment

Model	"LL"	Δdf	-2(ΔLL)	*p*-value	AIC	BIC
Comparison model	-162059.7	-	-	-	324217.4	324280.3
Base model	-166434.5	-	-	-	333063.0	333187.5
LEX predictor	-162592.9	4.0	7683.2	< 0.001	325387.8	325517.5
NP predictor	-166145.7	4.0	577.6	< 0.001	332493.4	332623.1
RC predictor	-165945.0	4.0	979.0	< 0.001	332092.0	332221.7
LEX + RC predictors	-160489.8	8.0	11889.4	< 0.001	321189.6	321324.4
All predictors	-160455.0	12.0	11959.0	< 0.001	321128.0	321267.9

Note: Δdf, -2(ΔLL), and *p*-value are for the omnibus tests between the base model and LC predictor models.

Table F14 presents the base model's item difficulties for the reference (OTH) and focal (SPA) groups and the differences in the item difficulties between groups. Table G28 presents the Rasch HGLM results for the base model and LC predictor models; the adjusted DIF estimates and confidence intervals are in Table G29 and the covariance matrix for the all predictors model is in Table G30. The covariance between lexical complexity and the intercept is large and positive, between complex noun phrases and the intercept is moderate and positive, between complex noun phrases and lexical complexity is small and negative, between relative clauses and the intercept is large and negative, between relative clauses and lexical complexity is large and negative, and between relative clauses and relative clauses is moderate and negative. In the base and RC predictor models, non-Spanish-speaking EBs had significantly higher abilities than

Spanish-speaking EBs; in the LEX and NP predictor models, there were no significant group differences in ability. In addition, there were no significant interactions between LC predictors and non-Spanish-speaking status.

Generally, items did not change DIF significance or direction between the base model and any LC predictor model. In the base model, b23 and b32 exhibited moderate DIF favoring Spanish-speaking EBs; after accounting for any LC predictors however, these items no longer exhibited significant DIF. Item b45 exhibited substantial DIF favoring Spanish-speaking EBs in the base model, but non-significant DIF after accounting for lexical complexity, complex noun phrases, or all LC predictors, although b45 exhibited substantial DIF favoring non-Spanish-speaking EBs after accounting for relative clauses.

Discussion

In this section, the results of the analyses for each comparison group will be summarized and discussed in relation to the three hypotheses for the study for both assessments. The three hypotheses were as follows:

1. LC factor scores will have significant main effects and interactions with emergent bilingual status; the interactions will favor English proficient students.
2. For items with higher LC, there will be less items flagged as significantly favoring EPs when including LC as a covariate.
3. For items with lower LC, there will be no change in items flagged as significantly favoring EPs when including LC as a covariate.

Hypothesis 1.

For the mathematics assessment, complex noun phrases factor scores had significant main effects and interactions with focal group status for the EP versus EB comparison groups.

The positive significant main effects for complex noun phrases indicate items with higher complex noun phrases factors scores are associated with increased ability estimates than items with lower complex noun phrases factor scores. These interactions did favor EPs; items with higher complex noun phrases factor scores were easier (lower item difficulty) for EPs than for EBs, although many items with high complex noun phrases factor scores exhibited DIF favoring EBs. The same patterns were found for the biology assessment for the LEX and NP predictor models (including exhibiting DIF favoring EBs for high factor scores), for all EP versus EB comparison groups except for one LC predictor model. For the NP predictor model for EPvLTEB, the interaction between complex noun phrases factor scores and LTEB status was not significant (p = .065). However, for the RC predictor model, relative clauses factor scores had significant interactions with focal group status across EP versus EB comparison groups, but these interactions favored EBs; items with higher relative clauses factor scores were easier for EBs than for EPs, although many items with high relative clauses factor scores exhibited DIF favoring EPs. The main effects of the LC factor scores LEX and NP predictor models for all EP versus EB comparison groups indicate items with higher lexical complexity or complex noun phrases factor scores are associated with increased ability estimates than items with low lexical complexity or complex noun phrases factor scores, while the main effects of the LC factor scores for the RC predictor model indicates items with lower relative clauses factor scores are associated with increased ability estimates than items with high relative clauses factor scores. These effects carry over into the all predictors models: items with higher lexical complexity are associated with increased ability estimates, after controlling for other factors; items with higher complex noun phrases are associated with increased ability estimates, after controlling for other factors; items with lower relative clauses are associated with increased ability estimates, after controlling for other factors.

In terms of significant interactions between focal group status and LC factor scores, the lexical complexity factor scores interactions with focal group status indicated items with higher lexical complexity factor scores were easier for EPs than for EBs, yet complex noun phrases factor scores interactions with focal group status were non-significant. There were mixed results when considering the interactions between relative clauses factor scores and focal group status. This interaction favored EBs for EPvEB, EPvSTEB, and EPvSPA, but the interaction was non-significant for EPvLTEB (p = .349) and EPvOTH (p = .215). These results suggest different LC features play different roles depending on the characteristics of the EBs taking the assessment. For the EB versus EB comparison groups, the main effects of each LC feature were significant in the single LC predictor models, but in the all predictors models, only relative clauses had a significant main effect; items with lower relative clauses factor scores were associated with increased ability estimates. For the single LC predictor models and all predictors models, there were no significant interactions between focal group status and any LC features.

Perhaps relative clauses are grammatical features test-takers rely on to identify information within items to successfully answer those items. While relative clauses do not necessarily contain the correct answer, the relative pronouns in relative clauses may “clue in” the test-taker to important information. Item b01 is reproduced below, with relative clauses underlined:

> The soybean aphid was introduced to the United States in 2000. The aphid killed many soybean plants. In 2004, scientists discovered <u>that some soybean plants were resistant to the aphid</u>. This resistance was genetically based. The scientists

> wanted to determine whether the resistant trait in these soybean plants has a dominant inheritance pattern.
>
> Which of the following would provide the best evidence <u>that the trait is dominant?</u>
>
> A. Two resistant plants are crossed, and none of the offspring are resistant.
>
> B. Two plants <u>that are not resistant</u> are crossed, and all of the offspring are resistant.
>
> C. A resistant plant and a plant <u>that is not resistant</u> are crossed, and all of the offspring are resistant.
>
> D. A resistant plant and a plant <u>that is not resistant</u> are crossed, and none of the offspring are resistant. (p. 464)

To answer this item correctly (answer C), the test-taker needs to identify which combinations of plants (whether they are resistant or not resistant) produce offspring with a dominant resistant trait. Much of the relevant information in the item to answer the question is embedded in the relative clauses in the text ("that some soybean plants were resistant…" and "that the trait is dominant") and the answers ("plant that is not resistant"). This can be compared to items without any relative clauses such as b07, where the item text contains all the information needed to answer the item without sorting through the text to identify relevant information to answer the item:

> Which two body systems carry signals from one part of the body to another part of the body?
>
> A. circulatory and nervous
>
> B. digestive and respiratory

C. excretory and circulatory

D. excretory and nervous (p. 466).

To answer this item correctly, the test-taker does not need to identify specific parts of the item that are most relevant to answering the item correctly, as all of the item text is relevant to the question, unlike b01, which introduces extraneous information for answering the item correctly; this is not to say the extraneous information is not construct-relevant, as b01 contains construct-relevant information matching a real-world context. Given the significant interaction between relative clauses and focal group status in the all predictors model for some EP versus EB comparisons, some subgroups of EBs (EBs overall, STEBs, and Spanish-speaking EBs) may overall be using relative clauses to identify the text necessary to answering the item correctly, compared to EPs.

For the EP versus EB comparison groups, there is a possible explanation for why complex noun phrases have a significant interaction with focal group status in the NP predictor model, but not the all predictor models. Complex noun phrases were counted when a noun had multiple determiners, adjectives, and prepositional phrases that add complexity; this grammatical feature may have some overlap with lexical complexity, which is derived from word count, general academic vocabulary, and words with seven or more letters. Specifically, word count contributes to both features. While lexical complexity and complex noun phrases are distinct enough to have their own main effects, how focal group status interacts with these features may similar, as the increased word count may be contributing to both LC features, with lexical complexity serving as a stronger predictor interacting with focal group status as it directly measures word count rather than indirectly like complex noun phrases. A similar explanation may be applied to why relative clause interactions with focal group status are different between subgroups in the all predictor models, with the increased word count associated with relative clauses (these clauses provide additional descriptive information about the subject) masked by lexical complexity, a construct that specifically considers word count in an item.

Hypotheses 2.

Items changing DIF significance or direction were evaluated for items with high LC factor scores; LC factor scores were considered "high" in the single LC predictor models if the factor score was greater than one standard deviation above the mean. In the all predictors models, LC factor scores were considered "high" in the all predictors models if two or more factors had a high factor score and no low factor scores. For the mathematics assessment, items with high complex noun phrases generally exhibited significant DIF favoring EBs when complex noun phrases were accounted for, with the exception of the EPvOTH comparison group, where these items exhibited non-significant DIF. For the biology assessment, items with high lexical complexity exhibited significant DIF favoring EBs after accounting for lexical complexity, items with high complex noun phrases exhibited non-significant DIF after accounting for complex noun phrases, items with high relative clauses exhibited non-significant DIF after accounting for relative clauses, and items with two or more high factor scores exhibited non-significant DIF after accounting for all LC predictors. These results were consistent across different EP versus EB comparison groups, although there were differences between subgroups for which items were flagged as having significant DIF in the base model. Overall, partial support was found for this hypothesis. For items with high LC factor scores, it appears that different LC features have different effects on item difficulties for EBs, with accounting for lexical complexity leading to items with high factor scores in these features exhibiting DIF favoring EBs, and accounting for complex noun phrases, relative clauses, or all predictors leading to items exhibiting non-significant DIF.

Hypothesis 3.

Items changing DIF significance or direction were evaluated for items with low LC factor scores; LC factor scores were considered "low" in the single LC predictor if the factor score was more than one standard deviation below the mean for lexical complexity factor scores, and if the factor score was the lowest factor score value for complex noun phrases and relative clauses factor scores. LC factor scores were considered "low" in the all predictors models if two or more factors had a low factor score and no high factor scores. For the mathematics assessment, items with high complex noun phrases generally exhibited significant DIF favoring EBs when complex noun phrases were accounted for, with the exception of the EPvOTH comparison group, where these items exhibited non-significant DIF. For the biology assessment, items with low lexical complexity exhibited significant DIF favoring EBs after accounting for lexical complexity, items with low complex noun phrases exhibited non-significant DIF after accounting for complex noun phrases, items with low relative clauses generally exhibited non-significant DIF after accounting for relative clauses, although some items that favored EBs in the base model favored EPs in the RC predictor model. For the all predictors model, items with two more low factor scores were split between exhibiting significant DIF favoring EBs or non-significant DIF after accounting for all LC predictors. However, there appeared to be some subgroup differences for the EPvSPA comparison groups, items with two or more low factor scores exhibited significantly DIF favoring EBs after accounting for all LC predictors. As there were not many items favoring EPs in the base model for either assessment, Hypothesis 3 could not be answered directly, although insights were found on whether items with low LC factor scores changed DIF significance or direction.

Study Conclusions

Although MCAS assessments are designed as 3PL tests (with parameters for item difficulty, discrimination, and guessability), Rasch models were used in the present study to handle computational challenges in examining the effect of LC on item responses. Rasch modeling assumes equal discrimination values across all items, meaning all items have equal weight in determining ability estimates and discriminate between higher and lower ability test-takers similarly. However, including the discrimination parameter assumes items have different weights in determining ability estimates; items with lower discrimination parameters contribute less to person ability estimates, but in a Rasch framework it is assumed all items contribute equally to person ability estimates. This is a limitation in the present study, as the different weights of items were not examined, although future research could examine the how LC features in items may contribute to a discrimination parameter. The LC in items likely influences item discrimination parameters and could explain sources of bias in non-uniform DIF.

Despite this, I could still make inferences about the effect of linguistic complexity of the item responses of students from these groups for research purposes, although these models should not be used to make decisions about the individuals tested. Another limitation for this dissertation study is that it does not consider an EB's individual English proficiency in predicting the effects of LC on item responses. Future studies can incorporate individual English proficiency as a person property in EIRM.

Model fit was not improved between the comparison model and the base model, which indicates that the inclusion of focal group by item interactions, or DIF estimates, did not significantly improve model fit. However, the method in the present study includes all focal group by item interactions in the model, regardless of DIF significance. This leads to non-significant parameters included in the base model, which decreases model fit compared to the comparison model. The lack of improvement in model fit is also likely indicative of how items in the MCAS are likely bias-free, given that it is a thoroughly vetted state achievement test designed to be high-stakes inferences about the abilities of the students taking the assessment. Regardless, DIF needs to be examined and evaluated in assessments like this, as the presence of items with DIF indicates potential bias. Even if DIF doesn't improve model fit, especially using the HGLM DIF method which includes focal group by item interactions for all items, DIF should still be evaluated. The present study was intended to illustrate a method to evaluate how considering item covariates like LC can be used to identify potential sources of bias in items. The inclusion of these item covariates did significantly improve model fit compared to the base model, this indicates that accounting for LC predictors does improve model fit.

From the results of the present study, it can be concluded that accounting for LC found in assessment items influences item responses between EPs and subgroups of EBs, leading to differences in DIF direction and significance. Accounting for lexical complexity (biology only) and complex noun phrases in items led to significant DIF favoring EBs, while accounting for relative clauses (biology only) led to significant DIF favoring EPs in items with high relative clauses factor scores and significant DIF favoring EBs in items with low relative clauses factor scores. Future research might conduct think-a-louds with both EBs and EPs to see how they use information introduced by grammatical features to answer the item. Martiniello (2008) did a version of this study with Spanish-speaking EBs; items exhibiting DIF against EBs were presented to EBs in think-alouds and their responses were evaluated for the linguistic features the participants had difficulty interpreting. This study could be repeated with EPs to see what features they use to answer the items correctly, as well as other subgroups of EBs to determine if there are differences in how these items are interpreted.

Previous research examining the effect of lexical features on DIF between EBs and EPs has found somewhat consistent results that lexical features are correlated with DIF. Martiniello (2008) identified uncommon words in items exhibiting a high amount of DIF, and DIF against EBs was found to be significantly correlated with general academic vocabulary (Haag et al., 2013; Heppt et al., 2015) in some studies, but not in Kachchaf et al. (2016). Heppt et al. (2015) reported significant correlations between the number of words with more than three syllables and DIF against EBs. While there were not many items exhibiting DIF in the base model for the biology assessment, in the LEX predictor and all predictors models for EP versus EB comparison groups, after accounting for lexical complexity, items exhibiting DIF favored EBs.

Previous research examining the effect of noun phrases on DIF between EBs and EPs has found mixed results; the present study utilized the counting of complex noun phrases, noun phrases with the addition of combinations of determiners, modifiers, and prepositional phrases, to evaluate whether noun phrases with increased complexity are potential sources of DIF between EBs and EPs. One study found the number of noun phrases predicts DIF against EBs (Haag et al., 2013), but another study found no significant correlations between the number of noun phrases and DIF against EBs (Kachchaf et al. 2016). While there were not many items exhibiting DIF in the base model for the mathematics or biology assessments, in the NP predictor models for EP versus EB comparison groups, items exhibiting DIF favored EBs when items had average to high complex noun phrases factor scores. However, in the all LC predictors models for the biology assessment, the interaction between complex noun phrases factors scores and focal group status was not significant for the EP versus EB comparison groups. Examination of the covariance matrices in the all predictors models for the biology assessment reveals the covariances between complex noun phrases and the other factors as the smallest ones. As part of lexical complexity are lexical features for total number of words and general academic vocabulary – features that are also present in complex noun phrases which include combinations that increase word length and general academic vocabulary – complex noun phrases factor scores may have not been significant because the effects of complex noun phrases were instead accounted for by lexical complexity. This feature may be more indicative of lexical complexity than grammatical complexity.

Previous research examining the effect of relative clauses on DIF between EBs and EPs has found mixed results. Kachchaf et al. (2016) did not find significant correlations with DIF against EBs and relative clauses, and in Buono & Jang (2021), relative clauses were not a significant predictor of DIF. However, Loughran (2014) found relative clauses predicted uniform DIF against EBs for fourth graders and relative clauses predicted uniform DIF that favored EBs for eighth graders. While there were not many items exhibiting DIF in the base model for the biology assessment, accounting for relative clauses led to significant DIF favoring EPs in items with high relative clauses factor scores and significant DIF favoring EBs in items with low relative clauses factor scores. In the all LC predictors models for EP versus EB comparison groups, after conditioning for lexical complexity, complex noun phrases, and relative clauses, the only items that favored EPs were the ones with high relative clauses factors. Perhaps EPs, with their greater English proficiency, are able to use relative clauses in items more effectively to answer the item correctly. Future research might examine think-a-louds for both EBs and EPs to see if there are differences between these groups of students in what grammatical features they use to help them find answers in items, as they may different approaches based on their English proficiency.

Analyses were conducted with EB versus EB comparison groups to examine if there were group differences in how LC influences item difficulty for direct comparisons of EB subgroups. For the mathematics assessment, all LC predictors had significant interactions with focal group status except for the NP predictor model for STEBvLTEB. However, for the biology assessment, no LC predictors had significant interactions with focal group status for the STEBvLTEB or OTHvSPA, although there were group differences in ability in the base model for OTHvSPA. Perhaps this is because the biology assessment was more linguistically complex than the math assessment (in terms of number of LC features counted, standardized scores were used for this study within each subject derived from the factor models in Chapter Two; different models were created for each subject) and this increased LC may have effected subgroups of EBs differently. The increased LC in the biology assessment could lead to no differences in how LC influences item difficulty for all subgroups of EBs.

Wolf and Leon (2009) proposed examining EB students' opportunity to learn to evaluate whether EBs are introduced and taught about academic language appearing on assessments. If EBs have different opportunities to learn based on their subgroup characteristics, this may lead to difference in item responses on assessments. Perhaps STEBs and non-Spanish-speaking EBs may have different opportunities to learn mathematics content assessed in Massachusetts schools compared to LTEBs and Spanish-speaking EBs. LTEBs and Spanish-speaking EBs have higher rates of IEPs and homelessness than STEBs and non-Spanish-speaking EBs, respectively. These are factors that would influence access to taught content.

There is a need for conducting DIF analyses for subgroups for heterogenous populations like EBs and other minoritized populations such as students with disabilities. Although DIF analyses require large sample sizes for accurate results, DIF analyses between subgroups of EBs and EPs need to be conducted in order to make valid interpretations about the abilities of EBs so item response differences of EBs from smaller subgroups are not masked by EBs from larger subgroups (Faulkner-Bond & Sireci, 2015; Lane & Leventhal, 2015; Sirecei et al., 2018).

CHAPTER FOUR

Conclusion

This chapter will discuss overall findings from Study One and Study Two as they relate to the research questions posed in Chapter One. Afterwards, the study contributions and limitations are discussed, followed by recommendations for future research. This dissertation explored whether unnecessary linguistic complexity (LC) in mathematics and biology assessment items changes the direction and significance of differential item functioning (DIF) between subgroups of emergent bilinguals (EBs) and English proficient students (EPs). Due to inconsistencies in measuring LC in items, Study One adapted a rubric to count construct-irrelevant instances of specific grammatical features (passive voice, complex verbs, subordinate clauses, relative clauses, and complex noun phrases) in items and introduced a method for evaluating lexical features (total words, general academic vocabulary, words with seven or more letters) in items. The items were drawn from four content assessments administered to Massachusetts high school students: two biology assessments and two mathematics assessments. The consistency of raters' counts of grammatical features was evaluated with generalizability theory. These counts of grammatical and lexical features were modeled in factor analyses to evaluate the multidimensionality of LC and subsequent fit of multidimensional LC models. Factor scores obtained from the measurement models for lexical complexity, relative clauses, and complex noun phrases created in Study One were used for Study Two.

In Study Two, Rasch hierarchical generalized linear models (HGLMs) were created to evaluate DIF between different subgroups of EBs and EPs on a biology assessment and a mathematics assessment, as including LC as an item covariate may predict item responses differently by comparison group. Seven comparison groups were evaluated across two assessments (mathematics and biology): EPs versus EBs, EPs versus short-term EBs, EPs versus long-term EBs, short-term EBs versus long-term EBs, EPs versus Spanish-speaking EBs, EPs versus non-Spanish-speaking EBs, and non-Spanish-speaking EBs versus Spanish-speaking EBs (reference group versus focal group, respectively). For each comparison group, at least five models were created: a comparison model with all participants in the comparison group with the main effect of focal group status, a "base model" that evaluated DIF for the comparison groups with no LC item covariates, a model including lexical complexity as an item covariate ("LEX predictor"), a model including complex noun phrases as an item covariate ("NP predictor"), and a model including relative clauses as an item covariate ("RC predictor"). If LC predictor models improved model fit, models with multiple LC predictors were created.

While the base model did not significantly improve model fit compared to the comparison model for both assessments and all comparison groups, the base model was still used as items must be screened for DIF in order to ensure we are making valid and fair assessments for students from historically underrepresented populations like EBs. For the EP versus EB comparison groups on the mathematics assessment, model fit only improved with the NP predictor model, while the LEX, NP, and RC predictor models improved model fit for the EB versus EB comparison groups; a model with all LC predictors improved model fit for the EB versus EB comparison groups. For the biology assessment, the LEX, NP, and RC predictor models improved model fit for all comparison groups; a model with all LC predictors improved model fit for all comparison groups.

The main effects of the item covariates (LC factor scores) and their interactions with focal group status were evaluated, as were the number of items within a comparison group that had changes in DIF significance or direction when including a LC predictor. All LC predictors had consistent main effects across comparison groups. For the mathematics assessment, items with higher complex noun phrases factor scores were consistently associated with increased ability estimates for all comparison groups (NP predictor model), and items with higher lexical complexity (LEX predictor model, all predictors model) or relative clauses factor scores (RC predictor model, all predictors model) were consistently associated with increased ability estimates for all EB versus EB comparison groups. For the biology assessment and all comparison groups, items with higher lexical complexity (LEX predictor model, all predictors model) or complex noun phrases factor scores (NP predictor model, all predictors model) were consistently associated with increased ability estimates, and items with lower relative clauses factor scores (RC predictor model, all predictors model) were consistently associated with increased ability estimates, with one exception. In the all predictors models for the EB versus EB comparison groups, only relative clauses had a significant main effect.

There were some changes in interactions with LC predictors and focal group status. For the mathematics assessment and EP versus EB comparison groups, complex noun phrases interactions favored EPs. For the mathematics assessment and EB versus EB comparison groups, generally the interactions in the single LC predictor models generally favored STEBs compared to LTEBs and non-Spanish-speaking EBs compared to Spanish-speaking EBs, but when all LC predictors were included, no interactions between LC predictor and focal group status were significant. For the biology assessment and EP versus EB comparison groups, lexical complexity and complex noun phrases factor scores interactions generally favored EPs, and relative clauses factor scores interactions favored EBs and EB subgroups. For the biology assessment and EB versus EB comparison groups, regardless of whether examining the single LC predictor or all predictors models, no interactions between focal group status and LC predictor were significant.

Changes in DIF significance and direction were compared between the base model and LC predictor models for all comparison groups. For the mathematics assessment and EP versus EB comparison groups, after conditioning on complex noun phrases, items generally exhibited significant DIF favoring EBs, regardless of whether the complex noun phrases factor scores were high (one standard deviation above the mean) or low (due to floor effects, the lowest complex noun phrases factor score). For the EPvOTH comparison group, more items exhibited non-significant DIF than other comparison groups, but most items followed this pattern of exhibiting significant DIF favoring EBs after conditioning on complex noun phrases. For the mathematics assessment and EB versus EB comparison groups, results were mixed. For STEBvLTEB, no items exhibited significant DIF in the base model or in any of the LC predictor models. For OTHvSPA, most items exhibited non-significant DIF in the base model and the LC predictor models, but some items favored Spanish-speaking EBs in the base model. Interestingly, the items with high LC factor scores for OTHvSPA generally exhibited non-significant DIF after accounting for any LC predictors, while the items with low LC factor scores remained exhibiting significant DIF favoring EBs. Similar results were found for the biology assessment for the EB versus EB comparison groups, although items with low factor scores exhibited non-significant DIF for OTHvSPA.

For the biology assessment, for EP versus EB comparison groups, different changes in DIF direction and significance occurred depending on what LC predictors were included in the model. After conditioning on lexical complexity, items exhibited significant DIF favoring EBs, regardless of whether the lexical complexity factor scores were high or low. After conditioning on complex noun phrases, items generally exhibited non-significant DIF regardless of whether the complex noun phrases factor scores were high or low, although some items that favored EBs in the base model continued to exhibit significant DIF favoring EBs in the NP predictor model. After conditioning on relative clauses, items generally exhibited non-significant DIF, however some items that favored EBs in the base model changed DIF direction to significantly favor EPs in the RC predictor model. Items with high relative clauses factor scores exhibited non-significant DIF after accounting for relative clauses, but items with low relative clauses factor scores exhibited significant DIF favoring EPs or non-significant DIF after accounting for relative clauses. Items with a low relative clauses factor score did not contain any relative clauses. After conditioning on all LC predictors, items were mixed on whether they exhibited non-significant DIF or significant DIF favoring EBs. For most EP versus EB comparison groups, about two-thirds of the items exhibited DIF favoring EBs, but for EPvOTH, about one-third of the items exhibited DIF favoring EBs, although this may be due to there being more items exhibiting non-significant DIF in the base model than other comparison groups. In the all predictors model, items were considered to have high factor scores when two or more predictors had high factor scores and no predictor had a low factor score, and items were considered to have low factor scores when two or more predictors had low factor scores and no predictor had a high factor score. Items with high factor scores in the all predictor models exhibited non-significant DIF when accounting for all LC predictors, but items with low factor scores were split between exhibiting non-significant DIF or significant DIF favoring EBs. These results indicate that the LC in items is not contributing to bias against EBs and may even be working in favor of EBs.

Items were less difficult for EBs than EPs after accounting for lexical complexity or complex noun phrases, which suggests the abilities of EBs are underestimated due to these features in items. Interestingly, items with low relative clauses factor scores favored EPs after accounting for relative clauses. Given the significant interaction between focal group status and relative clauses favoring EBs for some EP versus EB subgroup comparisons, the relative clauses in these items may have helped EBs interpret what information in the item needs to be used to answer the item correctly (in contrast to no relative clauses in an item), which suggests the complexity introduced by relative clauses is not detrimental to EBs. Items were less difficult for EBs than EPs after accounting for LC features, which suggests the abilities of EBs are underestimated due to LC in items, even if the items have low LC. Considering subgroup differences in these EIRMs, the key takeaway is that while different items are flagged as exhibiting significant DIF for different EP versus EB comparison groups when examining DIF with no LC predictors (base model), there are no subgroup differences in items changing DIF significance or direction after accounting for LC predictors.

Revisiting Research Questions

To determine how construct-irrelevant linguistic complexity in content assessment items influences the item responses of emergent bilinguals, five research questions were posited in Chapter One. Study One addressed the first two questions and Study Two addressed the remaining three questions. This section will revisit each research question as they relate to my findings.

Research Question 1

"How many raters are needed to reliably estimate the presence of five grammatical features in assessment items?"

To address this question, a generalizability theory decision study was conducted to evaluate how consistently four raters could count five grammatical features: passive voice, complex verbs, subordinate clauses, relative clauses, and complex noun phrases. Researchers have studied passive voice (Buono & Jang, 2021; Banks et al., 2016; Matiniello, 2008), complex verbs (Martiniello, 2008; Shaftel et al., 2006), subordinate clauses (Buono & Jang, 2021; Banket et al., 2016; Kachchaf et al., 2016), relative clauses (Buono & Jang, 2021; Banket et al., 2016; Kachchaf et al., 2016; Loughran, 2014), and complex noun phrases (Buono & Jang, 2021; Kachcaf et al.,2016; Heppt et al., 2015; Haag et al., 2013; Martiniello, 2008) because the grammatical complexity introduced by these features in assessment items may unfairly influence the responses of emergent bilingual test-takers. These features may influence the responses of students with disabilities as well; Abedi et al. (2010) adapted Shaftel et al.'s (2006) Linguistic Complexity Checklist into coding forms and guidelines for counting instances of grammatical features in assessments for the purpose of determining how these features influence the performance of students with disabilities on content assessments. Specifically, the rubric evaluates the cognitive, grammatical, lexical, and textual/visual features of the items; these dimensions were empirically supported with factor analysis. Part of Abedi et al.'s study examined the reliability of counts of grammatical features with coefficient α; these coefficient alphas ranged from .69 for counting relative clauses to .91 for counting complex verbs. Lexical and grammatical features were adapted from Shaftel et al. (2006) and raters were trained systematically to achieve acceptable reliability using coefficient α. Abedi et al.'s rubric was adapted for the present study (Appendix A) and used to train four raters (including the author) how to count grammatical features.

The raters used in the present study were graduate students in education with self-identified native or near-native proficiency in English. After raters were trained to count the five grammatical features, raters were given four assessments from the Massachusetts Comprehensive Assessment System (MCAS) and were asked to count the instances each feature appeared in each item. Two high school biology assessments (45 items each) and two high school mathematics assessments (42 items each) were used. This dissertation sought to evaluate how construct-irrelevant LC in items may lead to DIF between EBs and EPs. Construct-irrelevant LC specifically needs to be examined as construct-relevant vocabulary used on assessments is a construct intended to be measured by the instrument (Avenia-Tapper & Llosa, 2015). In order to evaluate whether LC is a potential source of bias leading to DIF against EBs, the LC accounted for must be construct-irrelevant. Therefore, raters were also asked to provide the construct-relevant count of grammatical features in items by counting the number of times a feature included construct-relevant vocabulary based on a provided wordlist of construct-relevant vocabulary for each subject (Appendices B & C). Construct-relevant counts were subtracted from total counts to obtain construct-irrelevant counts.

A multivariate single-facet decision study was conducted with items fully crossed with raters, and items and raters crossed with grammatical features, in order to evaluate the number of raters required to consistently count construct-irrelevant grammatical features in items. Generalizability and dependability coefficients (ρ_f^2 and ϕ_f, respectively) were calculated for each feature for two, three, four, five, and six raters; coefficients were considered sufficiently reliable above .800, assuming 90 items across the two biology assessments and 84 items across the two mathematics assessments. It is important to have an accurate average rating across items due to the difficulty in obtaining reliable estimates for counting grammatical features. Even with content experts rating the complexity of or counting grammatical features, the coefficient α or intraclass correlations are inconsistent. Haag et al.'s (2013) and Heppt et al.'s (2015) studies that examined the count of linguistic features used two-way random effects models to calculate intraclass correlation coefficients and reported the range of intraclass correlation coefficients. Haag et al. reported their coefficients ranged from .79 for counting noun phrases to 1.00 for counting total number of words and Heppt et al. reported their coefficients ranged from .75 for counting academic vocabulary (general and specialized) and 1.00 for counting total number of words, sentences, and words with at least three syllables. Instead of having their raters count or individual features, Lee and Randall (2011) rated items on their lexical and grammatical complexity holistically by having raters rate the items on a scale of one to five. The resulting intraclass correlation coefficients were .31 for lexical complexity ratings and .42 for grammatical complexity ratings. Given this range in accuracy of counts and ratings in past studies and the low consistency of counts of grammatical features in the present study, many items should be counted and rated to obtain reliable counts of grammatical features. The results of the present study also suggest the grammatical features in an assessment have varying consistency depending on the subject area of the items; construct-irrelevant counts of passive voice, complex verbs, and subordinate clauses were much less consistent for the mathematics assessments than the biology assessments (Table 2.3).

For the total count of grammatical features on the mathematics assessments, six raters would consistently count passive voice, four raters would consistently count relative clauses, and two raters would consistently count complex noun phrases. Six raters would not be enough to consistently count total instances of complex verbs or subordinate clauses on the mathematics assessments. For the construct-irrelevant count of grammatical features on the mathematics assessments, four raters would consistently count relative clauses, and three raters would consistently count complex noun phrases. Six raters would not be enough to consistently count construct-irrelevant instances of passive voice, complex verbs, or subordinate clauses on the mathematics assessments.

For the total count of grammatical features on the biology assessments, four raters would consistently count passive voice, five raters would consistently count complex verbs, five raters would consistently count relative clauses, and three raters would consistently count complex noun phrases. Six raters would not be enough to consistently count total instances of subordinate clauses on the biology assessments. For the construct-irrelevant count of grammatical features on the biology assessments, six raters would consistently count complex verbs, and five raters would consistently count complex noun phrases. Six raters would not be enough to consistently count construct-irrelevant instances of passive voice, subordinate clauses, or relative clauses on the biology assessments, although passive voice and relative clauses were close to meeting the threshold for consistency with six raters.

Raters were considerably less consistent in the construct-irrelevant counts compared to the total counts of grammatical features. While some grammatical features can be coded consistently by raters, identifying whether these features contain construct-irrelevant vocabulary is more difficult. The lack of consistency in counting grammatical features on these assessments suggests there is a need for better training of raters so features are not under-counted and also the need for recruiting content experts; this will be discussed further when considering study limitations. There were also less grammatical features in the mathematics assessments compared to the biology assessments when looking at the grand mean of raters' counts (Table 2.1). If raters are under-counting features, this would influence the consistency of counts of grammatical features in mathematics assessments more.

In addition, some grammatical features, like passive voice, complex verbs and subordinate clauses may be more difficult to rate overall. For passive voice, raters needed to be able to identify reduced passive voice, which may be more difficult to detect; for complex verbs, raters had to be able to identify many different complex verb structures including varying auxiliaries such as present participles, infinitives, and modals; for subjective clauses, raters need to be able to identify implied subordinate conjunctions such as “that.” Relative clauses and complex noun phrases on the other hand had less of these implied words or conjunctions (in the case of passive voice and subordinate clauses) and complicated rules (in the case of complex verbs). Raters may have had an easier time identifying relative clauses and complex noun phrases because of more salient features in these items, such as a relative pronoun at the beginning of the clause, and consistent propositional phrases.

Research Question 2

“What contributions do lexical features make to a lexical complexity factor score? What contributions do grammatical features make to a grammatical complexity factor score? What contributions do lexical complexity and grammatical complexity factors make to a LC factor score? Is LC measured this way multidimensional?”

To address this question, the construct-irrelevant grammatical feature counts from the generalizability theory decision study and lexical feature counted using total words in an item, count of unique general academic vocabulary in an item, and number of words with seven or more letters in an item were analyzed via factor analysis. Factor scores from this analysis were used as item covariates for the DIF analyses conducted in Study Two.

First, the unidimensionality of LC was tested before determining what contributions counts of features made to factor scores. A unidimensional model will all observed indicators (counts of lexical and grammatical features) loading onto one factor for LC was tested for each subject. This model’s fit statistics were then compared to those of a corresponding multidimensional model with all observed indicators loading onto their specific features’ factors (e.g., lexical features load onto a lexical complexity factor, passive voice counts for each rater load onto a passive voice factor, etc.). If the multidimensional model is better fitting than the unidimensional model, then LC is multidimensional for that subject. Due to issues with consistency in raters' counts of some grammatical features, multiple multidimensional models omitting those features were explored for both subjects. Three dimensional models were selected as the best-fitting multidimensional models, with factors providing sufficient internal consistency evidence for lexical complexity, relative clauses, and complex noun phrases.

The original factor analysis plan was to test whether higher-order models fit the data better than the multidimensional models. To establish a composite of LC, a model with the relative clauses, complex noun phrases, and lexical complexity factors loading onto an LC factor must fit better than a multidimensional model. To establish a composite of grammatical complexity, a model with the relative clauses and complex noun phrases factors loading onto a grammatical complexity factor must fit better than a multidimensional model (this multidimensional model would not include lexical complexity). However, due to only having three factors, the fit of models with a higher-order LC factor could not be tested. Similarly, due to only having two factors for grammatical features, the fit of models with a higher-order grammatical complexity factor could not be tested. Therefore, the multidimensional models (with measurement model variations as described in Study One) were the most appropriate and best-fitting models for counts of linguistic features for both subjects.

Measurement models were created for each factor (lexical complexity, relative clauses, and complex noun phrases) for each subject. Factor scores were extracted from these measurement models for use in Study Two as item covariates in DIF analyses. For both subjects on the lexical complexity factor, total words and number of words with seven or more letters had larger factor loadings than unique counts of general academic vocabulary.

As a higher-order factor for grammatical complexity could not be evaluated, specific features contributions to a grammatical complexity factor could not be evaluated. Raters' factor loadings and residual variances were examined for relative clauses and complex noun phrases. The higher a factor loading, the lower the residual variance was, with three of the four raters for relative clauses and complex noun phrases having high factor loadings and low residual variances For relative clauses on the mathematics assessments, one rater's counts had to be omitted because there was no residual variance (i.e., counted no instances of construct-irrelevant relative clauses). This rater consistently had lower counts of relative clauses and complex noun phrases than other raters, which led to low factor loadings and high residual variances for complex noun phrases on the mathematics and biology assessments and relative clauses on the biology assessments. This suggests the need for improved training or utilizing content experts as raters, which would lead to more consistent counts. If more consistent counts of grammatical features can be obtained, future research can examine the multidimensionality of grammatical complexity in assessment items.

Research Question 3

"How does linguistic complexity of the test item affect item difficulty for EBs compared to non-EBs on content assessments?"

To address this research question, in Study Two, the main effects of LC predictors and interactions between EB status and LC predictors (lexical complexity, relative clauses, and complex noun phrases) on item responses were evaluated for multiple comparison groups. Comparison groups for DIF analyses are presented in Table 4.1 (identical to Table 3.7, the first group listed for each comparison group is the reference group and the second group listed is the focal group). The EP versus EB comparison groups (EPvEB, EPvSTEB, EPvLTEB, EPvSPA, and EPvOTH) are discussed in research questions 3-5, with the EB versus EB comparison groups' results (STEBvLTEB and OTHvSPA) presented in the "Additional Findings" section, as these three research questions were concerned with EP versus EB comparisons.

Table 4.1.

Comparison Groups for DIF Analyses

Comparison Group Category	Groups Compared	Comparison Group Abbreviation
Baseline	EP vs. EB	EPvEB
Length of time as EB	EP vs. STEB EP vs. LTEB STEB vs. LTEB	EPvSTEB EPvLTEB STEBvLTEB
First language	EP vs. Spanish-speaking EB EP vs. Non-Spanish-speaking EB Spanish-speaking EB vs. Non-Spanish-speaking EB	EPvSPA EPvOTH OTHvSPA

For analyses, first, model fit between comparison models and models examining the effect of focal group status and DIF (base model) were compared for each group. Although model fit did not improve with the inclusion of DIF estimates (for all comparison groups), DIF analyses were conducted because items still need to be routinely screened for DIF to have valid scores for students from historically underrepresented groups, like EBs. Second, it needed to be determined if the inclusion of LC predictors improved model fit. Rasch HGLMs with LC factor scores as item covariates were created to evaluate DIF between subgroups of EBs and EPs on a biology assessment and a mathematics assessment. Separate HGLMs were estimated for each LC predictor and comparison group. Models with multiple LC predictors were created if single LC predictor models improved model fit compared to the base model.

In the mathematics assessment for the EP versus EB comparison groups, only complex noun phrases factor scores significantly improved model fit compared to a model without LC predictors; for this assessment, combinations of LC predictors were not explored as lexical complexity and relative clauses factor scores did not significantly improve model fit. In the biology assessment, all three LC predictors significantly improved model fit; models with all LC predictors were analyzed and for each EP versus EP comparison group, models with all LC predictors improved model fit.

For the mathematics assessment, after accounting for complex noun phrases factor scores as a predictor, the significant positive main effect of complex noun phrases indicated items with higher complex noun phrases factor scores were associated with increased ability estimates. However, the interactions between complex noun phrases and EB status were significant, indicating there are group differences in how complex noun phrases factor scores influence item responses. The positive interactions for the interaction between complex noun phrases factor scores and EB status indicated items with higher complex noun phrases factor scores were significantly easier for EPs than for EBs and subgroups of EBs, holding other variables constant.

For the biology assessment, after accounting for lexical complexity factor scores as a predictor, the positive main effect of lexical complexity indicated that items with higher lexical complexity scores were associated with increased ability estimates EBs had a significantly higher ability than EPs. The interactions between lexical complexity and EB status were significant, indicating there are group differences in how lexical complexity factor scores influence item responses after conditioning on overall item difficulty, level of lexical complexity, and EB status.

The positive interaction between lexical complexity factor scores and EB status indicated items with higher lexical complexity factor scores were significantly easier for EPs than for EBs and subgroups of EBs after conditioning on overall item difficulty, level of lexical complexity, and EB status. The same findings were found for the complex noun phrases predictor models for the EPvEB, EPvSTEB, EPvSPA, and EPvOTH comparison groups. However, for EPvLTEB, accounting for complex noun phrases led to no significant interactions between complex noun phrases factor scores and LTEB status, although complex noun phrases factor scores had a significant main effect. For all EP versus EB comparison groups, the significant positive main effect of complex noun phrases indicated items with higher complex noun phrases factor scores were associated with increased ability estimates. A different pattern emerged for the relative clauses predictor. After accounting for relative clauses factor scores as a predictor, the significant negative main effect of relative clauses indicated items with lower relative clauses factor scores were associated with increased ability estimates. The interactions between relative clauses and EB status were significant, indicating there are group difference in how relative clauses factor scores influence item responses after conditioning on overall item difficulty, level of relative clauses, and EB status. The negative interaction between relative clauses factor scores and EB status indicated items with higher relative clauses factor scores were significantly easier for EBs and subgroups of EBs than for EPs responses after conditioning on overall item difficulty, level of relative clauses, and EB status.

To illustrate the effects of LC predictors and interactions with focal group status for the all predictors models, Table 3.26 is repeated as Table 4.2. The results of including all LC predictors are the same for EPvEB, EPvSTEB, and EPvSPA; items with higher lexical complexity factor scores are easier for EPs, items with higher relative clauses factor scores are easier for EBs, and there are no group differences in how complex noun phrases influence item responses. Results are similar for EPvOTH, but there are no group differences in how relative clauses influence item responses. The EPvLTEB multiple LC predictors model did not include complex noun phrases; items with higher lexical complexity factor scores are easier for EPs and there are no group differences in how relative clauses influence item responses. Test developers need to consider if the LC in items is a construct that is intended to be measured by their assessments. If the LC in assessment items is construct-irrelevant, bias may be introduced by items that are unnecessarily linguistically complex.

Table 4.2.

Significance of LC Factor Predictors and Interactions with Focal Group Status for EP Versus EB Comparison Groups for Multiple LC Predictor Models – Biology Assessment

Comparison Group	LEX		NP		RC	
	Main effect	Interaction	Main effect	Interaction	Main effect	Interaction
EPvEB	***	Favors EPs ***	***	0.162	***	Favors EBs *
EPvSTEB	***	Favors EPs ***	***	0.265	***	Favors STEBs *
EPvLTEB	***	Favors EPs *	***	0.310	***	0.349
EPvSPA	***	Favors EPs ***	***	0.437	***	Favors SPAs *
EPvOTH	***	Favors EPs ***	***	0.169	***	0.215

Note: *** = $p < .001$, * = $p < .05$. If γ_{s1} was not significant, *p*-values were listed instead.

Research Question 4

“Does accounting for linguistic complexity lead to differences in uniform DIF significance or direction when evaluating DIF between EBs and non-EBs?”

Study Two addressed this question with two specific hypotheses. The first hypothesis was "For items with higher LC, there will be less items flagged as significantly favoring EPs when including LC as a covariate." Items with significant DIF and higher LC are expected to favor EPs per the results of Wolf and Leon (2009). When LC is accounted for, then items with higher LC that favor EPs should either exhibit non-significant DIF or favor EBs, as the presumed source of DIF is accounted for. By accounting for the effect of LC on item responses, if LC is influencing item responses, then there should be no DIF in items after accounting for LC. The second hypothesis was "For items with lower LC, there will be no change in items flagged as significantly favoring EPs when including LC as a covariate." Items with significant DIF favoring EPs and lower LC are not expected to change DIF direction or significance because the source of DIF (some factor that is not LC) was not accounted for. "Base models" without an item covariate for LC were compared to models including one of the three LC factors in order to evaluate which items changed DIF significance or direction when accounting for lexical complexity, relative clauses, or complex noun phrases. To determine the significance of the adjusted DIF estimates that took γ_{01} into account, 95% confidence intervals were calculated; if the confidence interval contained γ_{01}, the adjusted DIF estimate was not significant. ETS's procedure for classifying the magnitude of DIF was utilized (Zwick, 2012; Monahan, et al., 2007). By taking the odds-ratios of the item by focal group status interaction plus the group differences in item responses ($\gamma_{q1} + \gamma_{01}$) and using Equation 3.8, the magnitude of DIF can be interpreted for the base model (Monahan, et al., 2007). For the models including LC predictors, the odds-ratio of the sum of the item by focal group status interaction, group differences in item responses, and LC predictor by focal group interaction ($\gamma_{q1} + \gamma_{01} + \gamma_{s1}$) is used to determine the effect size of DIF. The analyses discussed in this section are for the EPvEB comparison group. Items were evaluated as having a high LC factor score if that score was greater than one standard deviation above the mean and a low LC factor score if that scores was greater than one standard deviation below the mean (lexical complexity) or the lowest LC factor score for that feature due to ceiling effects for some LC features (complex noun phrases and relative clauses). Items in the all predictor models were considered as having high LC factor scores if they had two or more high LC factor scores and no low LC factor scores, and as having low LC factor scores if they had two or more low LC factor scores and no high LC factor scores.

For the mathematics assessment, in the base model, items were split between exhibiting significant DIF favoring EBs (16 items), significant DIF favoring EPs (14 items), or non-significant DIF (11 items). However, most items exhibited significant DIF favoring EBs after accounting for complex noun phrases. In the NP predictor model, after accounting for complex noun phrases, all items favoring EBs or EPs in the base model exhibited significant DIF favoring EBs after accounting for complex noun phrases, and six items changed from exhibiting non-significant DIF in the base model to exhibiting significant DIF favoring EBs. Of the items with high complex noun phrases factor scores, one item exhibiting significant DIF favoring EPs in the base model exhibited DIF favoring EBs after accounting for complex noun phrases, and four items exhibiting DIF favoring EBs in the base model continued to exhibit DIF favoring EBs after accounting for complex noun phrases, although these four items were polytomous and in the present study, polytomous items tended to not change DIF significance or direction after accounting for LC features. Of the items with low complex noun phrases factor scores, five items exhibiting significant DIF favoring EPs in the base model exhibited DIF favoring EPs after accounting for complex noun phrases. Two other items exhibited significant DIF favoring EBs in the NP predictor model (one item exhibited non-significant DIF in the base model and another exhibited significant DIF favoring EBs in the base model), and one item exhibited non-significant DIF in both models. These results suggest for the mathematics assessment, the mathematics ability of EBs may be under-estimated on items with low complex noun phrases factor scores.

For the biology assessment, in the base model, items were split between exhibiting significant DIF favoring EBs (16 items), significant DIF favoring EPs (five items), or non-significant DIF (23 items). However, most items exhibited significant DIF favoring EBs after accounting for LC predictors. In the LEX predictor model, after accounting for lexical complexity, all items exhibited significant DIF favoring EBs. In the NP predictor model, after accounting for complex noun phrases, all items exhibited non-significant DIF, except for the five polytomous items which exhibited significant DIF favoring EBs in the base model and the NP predictor model. Due to all the items in the LEX and NP predictor models exhibiting the same type of DIF (significantly favoring EBs and non-significant DIF, respectively), the effect of high or low factor scores could not be evaluated for these factors. In the RC predictor model, after accounting for relative clauses, most items exhibited non-significant DIF except for the five polytomous items and three of the dichotomous items which exhibited significant DIF favoring EBs in the base model and significant DIF favoring EPs in the RC predictor model. Five of these items had low relative clauses factor scores, although the majority of items with low relative clauses factor scores exhibited non-significant DIF when relative clauses were accounted for. In the all predictors model, 32 items exhibited DIF favoring EBs and 12 items exhibited non-significant DIF after accounting for all LC predictors. Items with low factor scores were split between favoring EBs (6 items) or exhibiting non-significant DIF (4 items), and items with high factor scores, exhibited non-significant DIF. Taken all together, these results suggest for the biology assessment, the biology ability of EBs may be under-estimated due to lexical complexity, but not complex noun phrases or relative clauses.

The hypothesis that items with high LC factor scores would not favor EPs after accounting for that LC feature could not be directly answered as few items in the base model with high factor scores favored EPs. These items tended to favor EBs and exhibited non-significant DIF in the base model. Mild support was found for the hypothesis that items with low LC factor scores would not change DIF significance or direction. Many items with low LC factor scores exhibited non-significant DIF and continued to exhibit non-significant DIF in the LC predictor models, but in some models, items with low factor scores changed DIF significance or direction after accounting for LC predictors.

Test developers should consider the impact of accounting for LC on DIF, as after accounting for lexical complexity, complex noun phrases, relative clauses, multiple LC predictors, items tend to exhibit DIF favoring EBs, which suggests the ability estimates of EBs may be under-estimated. Whether the items have high or low factor scores for LC features compared to other items on the assessment also needs to be considered in the test development process, as items with a higher value of LC factor scores exhibit DIF favoring one group of test-takers over another. The inclusion of item covariates should be included in DIF analyses when there are item covariates that are consistently expected to be potential sources of DIF. In the case of EBs, unnecessary LC in items has been theorized to be a source of DIF between EBs and non-EBs by many researchers (Banks et al., 2016; Kachchaf et al., 2016; Abedi, 2015; Heppt, et al., 2015; Haag et al., 2013; Turkan & Liu, 2012; Lee & Randall, 2011; Sato et al., 2010; Wolf & Leon, 2009; Shaftel et al., 2006; Abedi & Lord, 2001). While revising items to contain less unnecessarily linguistically complex language is an important step, evaluating whether the LC in items changes the significance and direction of DIF will provide clearer evidence as to the effects of LC on assessment performance.

Research Question 5

"Which EB subgroups exhibit differential functioning? Are there differences by subgroups of EBs in how accounting for linguistic complexity affects uniform DIF significance or direction?"

To address this question, the analysis that was conducted for EBs and EPs was repeated for four additional comparison groups: EPvSTEB, EPvLTEB, EPvSPA, EPvOTH. For the EP versus EB subgroup comparison groups, generally the same patterns as EPvEB emerged, with some small differences. Generally, for all EP versus EB comparison groups, there were significant interactions with LC predictors and focal group status on both assessments, with the exception of the NP predictor model for EPvLTEB on the biology assessment. For EPvLTEB, this suggests that complex noun phrases may influence group differences in item difficulties on the mathematics assessment, but not the biology assessment. Generally, lexical complexity, complex noun phrases, and relative clauses influence group differences in item difficulties on both of these assessments between EBs and EPs, with lexical complexity and complex noun phrases decreasing item difficulty for EPs relative to EBs and relative clauses decreasing item difficulty for EBs relative to EPs.

Oliveri et al. (2014) concluded the heterogeneity of EBs and students with disabilities may lead to greatly reduced DIF detection rates, therefore it may be expected more items would be detected as having DIF for the EP versus EB subgroup comparison groups than for the EPs versus EBs group. However, based on the results of the present study, DIF detection rates were not reduced, but different items were detected for DIF based on subgroup characteristics. There were minor differences in changes in DIF significance or direction when comparing EP versus EB subgroup comparison groups to EPvEB; most of these changes had to do with whether an item exhibited significant DIF in the base model, which is likely attributable to type I error, although EPvOTH had less items detected as having DIF in the base model compared to the other EB versus EP comparison groups. Similarly, there were few subgroup differences in how items changed DIF significance or direction about accounting for LC features. Test developers should consider that while there may not be subgroup differences in how LC in items influence DIF detection, due to the presence of some subgroup differences in what items were identified as having DIF in the base model, DIF analyses based on subgroup characteristics may be warranted. Per the recommendations of Lane & Leventhal (2015), these DIF analyses need to become a routine part of evaluating items for bias, as subgroup characteristics that influence assessment performance may be masked.

Additional Findings

While not a research question, subgroups of EBs were compared to each other: STEBs to LTEBs and non-Spanish-speaking EBs to Spanish-speaking EBs. For the mathematics assessment, for STEBvLTEB and OTHvSPA, the main effects of LC features were significant and positive for the LEX, NP, RC predictor models; items with higher LC factor scores were associated with increased ability estimates, for all features. For the biology assessment, the main effects of LC features were significant and positive for the LEX and NP predictor models, but significant and negative for the RC predictor models; items with higher LC factor scores were associated with increased ability estimates for lexical complexity and complex noun phrases, but decreased ability estimates for relative clauses. This suggests there are differences in how relative clauses influence item difficulty between mathematics and biology. Perhaps in the biology assessment, which had more linguistic features by count than the mathematics assessment (Tables 2.1 & 2.10), relative clauses helped test-takers identify relevant information in the text to answer the item correctly, but relative clauses in the mathematics assessment did not, as these items generally contained less linguistic features.

In the all predictors models for the mathematics assessment, lexical complexity and complex noun phrases maintained significant positive main effects and relative clauses' main effect was non-significant for STEBvLTEB, and all three LC features maintained significant positive main effects for OTHvSPA. In the all predictors models for the biology assessment, only relative clauses maintained the significant negative main effect for both STEBvLTEB and OTHvSPA. These results suggest that when accounting for all three of these LC features, lexical complexity and complex noun phrases influence item difficulty for EBs in the mathematics assessment, but not the biology assessment, and relative clauses influences abilities estimates for EBs in the biology assessment. It is interesting that for the mathematics assessment, different results for the main effect of relative clauses appear for STEBvLTEB and OTHvSPA, as these comparison groups have the same sample, but after holding focal group status constant, different main effects for relative clauses emerge. The interactions between focal group status and LC features were examined for the single and all predictors models.

In the single predictor models for the mathematics assessments, the interactions between LC feature and focal group status tended to favor STEBs (STEBvLTEB) and non-Spanish-speaking EBs (OTHvSPA), except for the interaction between complex noun phrases and focal group status for STEBvLTEB. However, in the all predictors model, none of these interactions were significant. The interaction between relative clauses and focal group status was close to the threshold for significance for OTHvSPA (p = .082); this combined with the relative clauses main effect differences in the all predictor model between EB versus EB comparison groups suggests that there may be some differences in how relative clauses influence item responses after accounting for lexical complexity and complex noun phrases, further research may investigate this further. For the single and all predictor models for the biology assessment, the interactions between LC feature and focal group status were non-significant for STEBvLTEB and OTHvSPA. The differences between subjects may reflect some small differences in how LC features influences the item responses between EB subgroups, but these differences may not be large enough to consider as practically influencing item responses between EB subgroups.

Little DIF was detected between for EBs versus EBs comparison groups in the base models. For STEBvLTEB, no items exhibited significant DIF in the base model or any LC predictor model for both assessments. For OTHvSPA, few items exhibited significant DIF in the base model for either assessment; those that did exhibited significant DIF favoring Spanish-speaking EBs. Accounting for LC factor scores for OTHvSPA typically led to these items exhibiting non-significant DIF. These results suggest that little item bias exists between EB subgroups, and accounting for LC features minimizes what is present. However, test developers should consider examining think-a-louds with EBs from varying subgroups to determine if there are differences in how EBs think about and use the linguistic features in items.

Study Contributions

The present study demonstrates the need for consistent measurements of LC in items. The accuracy of ratings of linguistic features needs to be taken into consideration, as raters have difficulty consistently identifying specific grammatical features. In Study One, raters under-identified grammatical features, regardless of whether they were total counts or construct-irrelevant counts. Content experts are needed to count or rate linguistic features in items, and extensive training needs to be provided to raters. Despite this, complex noun phrases, relative clauses, and lexical features were consistently counted, and evidence was found for a multidimensional model of LC in assessment items. Although the present study could not establish a factor for grammatical complexity due to inconsistent rater counts of features, future research can accomplish this by having trained content experts serve as raters to obtain more consistent rater counts. Test developers can use the present study as a framework for counting linguistic features and utilizing factor analysis to identify factor scores for lexical and grammatical complexity for use as item covariates in IRT models to account for the effect of LC on item responses, as this is a potential source of bias in items influencing group differences in item responses between EBs and non-EBs. Test developers need to consider if the LC in items is a construct that is intended to be measured by their assessments. If the LC in assessment items is construct-irrelevant, bias may be introduced by items that are unnecessarily linguistically complex. By accounting for unnecessary LC in EIRMs, the person ability estimates of EBs will be less biased, leading to more accurate inferences about the mathematics and biology abilities of EBs; the results of the present study demonstrate the ability estimates of EBs may be underestimated because of unnecessary LC introduced into items.

EIRMs can be used to explore what item properties influence differences in item responses between groups. The present study used LC in items, but other item properties such as whether an item is multiple choice or free-response or subscales (e.g., geometry, statistics and probability, expressions and equations, etc.) can be included in EIRMs evaluating content assessment. By identifying whether there are group differences in item properties on content assessment, test developers can determine whether the language used in their assessments is intended to be measured and used to be inferences about test-takers. Otherwise, LC in items may need to be accounted for as a covariate when examining DIF between EBs and non-EBs in order to obtain more accurate ability estimates from test-takers. While revising items to contain less unnecessary LC is an important step to take in the test development process, evaluating whether accounting for the LC in items changes the significance and direction of DIF directly examines whether the source of DIF in items may be explained by LC.

In addition, this dissertation evaluated how dividing EBs into subgroups based on demographic characteristics (STEBs, LTEBs, Spanish-speaking EBs, and non-Spanish-speaking EBs) influences DIF detection. Lane and Leventhal (2015) argued different groups of EBs have different needs and if differences in their item responses are identified, this may indicate a need for instructional change or different considerations in assessing these subgroups. DIF analyses by subgroups of EBs are uncommon in assessment research due to the large sample sizes required, but there is certainly enough power to conduct DIF analyses for some subgroups at the state-level for content assessments. State Boards of Education should examine subgroups of historically underrepresented populations in their routine DIF analyses to determine if items are influencing some subgroups of these populations differently.

In Study Two, in the base models without LC predictors, some items were detected as having DIF for subgroups of EBs that were not detected in DIF analyses with all EBs. Similarly, some items were detected as not having DIF for EPvOTH that were detected in DIF analyses with all EBs. These results somewhat support and contract the findings in Oliveri et al.'s simulation study (2014) that the heterogeneity of EBs may lead to greatly reduced DIF detection rates. Based on the results of Study Two, DIF detection rates were not reduced, rather different items may be detected for DIF based on subgroup characteristics. Based on these results, state

Boards of Education should evaluate subgroups of EBs, and also students with disabilities (Lane & Leventhal, 2015).

Study Limitations

In this dissertation, lexical complexity, complex noun phrases, and relative clauses were found to be significant predictors of group differences in item responses between EBs and EPs in models with a single LC predictor. However, when accounting for multiple LC predictors in a model, changes in DIF significance and direction from a model without LC predictors most closely followed the model with only the lexical complexity predictor. This may be related to previous findings; due to inconsistencies in how linguistic features predict differences in item difficulties between EBs and EPs, LC should be evaluated as composites of lexical and grammatical complexity (Avenia-Tapper & Llosa, 2015; Lee & Randall, 2011; Martiniello, 2009). As lexical complexity was a composite of lexical features, this is what may have made lexical complexity a better predictor of LC in items compared to using the individual grammatical features of complex noun phrases and relative clauses. There are mixed results as to what aspects of English grammar are the most difficult for EBs to master, even between instructors and their students. In their study on the difficulty of English grammar features on Iranian undergraduates learning English, Dehghani et al. (2016) found relative clauses to be the least difficult grammatical feature based on participant performance, but instructors perceived relative clauses to be one of the most difficult grammatical features. There are inconsistencies between studies as to what grammatical features are even examined (Deghani et al., 2016; Shiu, 2011; Darus & Subraminian, 2009)

Multicollinearity is a possible issue in the present study, although steps were taken to reduce the effects of multicollinearity which included standardizing the counts of linguistic features used in factor analyses, standardizing the factor scores extracted from the factor analyses. While there were high, positive correlations between factors for the mathematics assessment (LEX and NP r = .714, LEX and RC r = .647, NP and RC r = .601), these factors were less highly correlated for the biology assessment LEX and NP r = .678, LEX and RC r = .454, NP and RC r = .522).

Due to raters' lack of consistency in rating passive voice, complex verbs, and subordinate clauses, a multidimensional model of grammatical complexity could not be established in Study One. This prevented a composite of grammatical complexity from being used a LC predictor in Study Two. Graduate students likely cannot be trained to count grammatical features consistently without extensive training, and content experts should be utilized instead. In Abedi et al.'s study with the same rubric, coefficient alphas ranged from .69 for counting relative clauses to .91 for counting complex verbs; in the present study, relative clauses were counted much more consistently than complex verbs. Other research that has reported the inter-rater reliability for counting or rating grammatical features tends to report ranges of reliability coefficients; it is unclear how reliably specific grammatical features are counted using these methods. Haag et al.'s and Heppt et al.'s studies that examined the count of linguistic features used two-way random effects models to calculate intraclass correlation coefficients and reported the range of intraclass correlation coefficients. Haag et al. reported their coefficients ranged from .79 for counting noun phrases to 1.00 for counting total number of words and Heppt et al. reported their coefficients ranged from .75 for counting academic vocabulary (general and specialized) and 1.00 for counting total number of words, sentences, and words with at least three syllables. Instead of having their raters count or individual features, Lee and Randall (2011) rated items on their lexical and grammatical complexity holistically by having raters rate the items on a scale of one to five. The resulting intraclass correlation coefficients were .31 for lexical complexity ratings and .42 for grammatical complexity ratings.

In terms of generalizability, the dataset used includes the item-level responses of all Massachusetts high school students that took the 2019 high school biology and 10th grade mathematics assessment and therefore results are generalizable to that population and those subjects, but may not be generalizable to subjects outside of science and mathematics, earlier grade levels, or to other states' content assessments. In addition, these assessments were designed to be used with 2PL, 3PL, and graded response models and this dissertation evaluated the assessments with Rasch rating scale HGLMs (1PL). Rasch modeling assumes equal discrimination values across all items, meaning all items have equal weight in determining ability estimates and discriminate between higher and lower ability test-takers similarly. However, including the discrimination parameter assumes items have different weights in determining ability estimates; items with lower discrimination parameters contribute less to person ability estimates, but in a Rasch framework it is assumed all items contribute equally to person ability estimates. LC features in items may contribute to the discrimination parameter, or how much weight an item contributes to person ability estimates.

Model fit did not improve between the HGLMs accounting for group differences in item responses ("comparison model") and the HGLMs accounting for group differences in item responses and focal group by item interactions (DIF estimates, "base model). This was due to the method being used which called for the inclusion of all focal group by item interactions to evaluate DIF for all items. This model also allowed the evaluation of which items changed DIF significance or direction after accounting for LC predictors. Although model fit did not improve when considering DIF in the model, this does not mean that items should not be screened for

DIF, even in assessments like the MCAS where items undergo a review process to ensure assessed students are exposed to items that are bias-free.

Despite these limitations, I still established consistent counts of some grammatical features, and obtained well-fitting factor scores for lexical complexity, complex noun phrases, and relative clauses for use in evaluating whether LC influences item responses between EBs and EPs. In addition, I found evidence that LC factor scores significantly interact with EB status to explain group differences in item responses on a biology and a mathematics assessment, although there were little to no subgroup differences in how LC factor scores influenced item responses or changes in DIF significance or direction. The consistency of counting grammatical features could be improved upon by using content experts to count these features; with this, well-fitting multidimensional models of grammatical complexity could be established to determine whether composites of grammatical complexity influence item responses between EBs and EPs.

Recommendations for Future Research

If more consistent counts of grammatical features can be obtained, future research can examine the presence of a higher-order grammatical complexity factor when counting grammatical features in assessment items. With at least four factors of grammatical complexity, the presence of a higher-order grammatical complexity model could be tested. Due to inconsistencies in research with what linguistic features influence item difficulty, LC needs to be scored as a composite of overall LC (Martiniello, 2009). Due to the multidimensionality of LC, other researchers have recommended looking at composites of lexical complexity and grammatical complexity (Avenia-Tapper & Llosa, 2015; Lee & Randall, 2011; Wolf & Leon, 2009).

Study Two could be replicated as with 2PL or 3PL models using the PROC NLMIXED procedure in SAS instead of a Rasch HGLM modeling approach. The LC in items likely influences item discrimination parameters and could explain sources of bias in non-uniform DIF. In addition, future research might conduct think-a-louds with both EBs and EPs to see how they use information introduced by grammatical features to answer the item. Martiniello (2008) did a version of this study with Spanish-speaking EBs; items exhibiting DIF against EBs were presented to EBs in think-alouds and their responses were evaluated for the linguistic features the participants had difficulty interpreting. This study could be repeated with EPs to see what features they use to answer the items correctly, as well as other subgroups of EBs to determine if there are differences in how these items are interpreted.

While LC has been found to be predictive of item difficulty, the contextual factors of EBs (such as native language or dialect spoken, language of assessment items) taking the test must be taken into consideration when evaluating item difficulty as it pertains to EBs (Solano-Flores, 2014). LC also affects EBs differently depending on these contextual factors. While LC was found in this dissertation to affect item difficulty, LC affected item difficulty differently depending on the EB subgroup examined. Solano-Flores (2014) suggests the LC of items and characteristics of the test-taking population needs to be considered during test development through expert reviewers at all stages of the test development process using experts from a variety of professional backgrounds ("e.g., teachers, sociolinguists, translators, content experts, test developers" p. 240). Test developers should include item covariates that are predicted sources of bias for historically under-represented populations in their DIF analyses, such as LC item covariates for DIF analyses between EBs and non-EBs, as not including these covariates may lead to the abilities of test-takers from the historically under-represented group being underestimated. In addition, EBs needed to be included throughout the test development process, and not only EBs from majority groups, such as Spanish-speaking EBs. Performance on and cognitive interviews with the items will provide different insights as to how items influence the responses of EBs. The relationship between LC and visual devices could also be examined, as the presence of visual representations on assessments such as charts, graphs, number lines, and Venn diagrams may influence how EBs interpret the language in items (Solano-Flores, et al., 2014).

Closing Summary

This dissertation examined specific linguistic features predicted to influence the item responses of EBs in content assessment. In previous studies examining the relationship between LC and DIF, it was unclear exactly how reliably linguistic features could be counted as a range of coefficients was typically reported, if it was reported at all. In Study One, I investigated whether graduate students could be trained to consistently count features by adapting a rubric from Abedi et al. (2011). I found that complex noun phrases and relative clauses could be counted consistently by this sample, but not passive voice, complex verbs, and subordinate clauses. In Study Two, I included LC into IRT models to explain potential sources of bias that may cause DIF in content assessments and found lexical complexity, complex noun phrases, and relative clauses to significantly influence group differences in item responses. This study is different from previous studies in that it includes LC as a covariate directly into the IRT model (Kachchaf et al., 2016; Heppt et al., 2015; Haag et al., 2013; Wolf & Leon, 2009). By including LC as item covariates in explanatory IRT models (EIRMs), potential sources of bias can be directly identified.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231-257. https://doi.org/10.1207/S15326977EA0803_02

Abedi, J. (2015). Language issues in item-development. In S. Lane, M. S. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed.). Florence, KY: Routledge. https://doi.org/10.4324/9780203102961-26

Abedi, J., Bayley, R., Ewers, N., Mundhenk, K., Leon, S., Kao, J., & Herman, J. (2012). Accessible reading assessments for students with disabilities. *International Journal of Disability, Development and Education*, *59*(1), 81–95. https://doi.org/10.1080/1034912X.2012.654965

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2010). Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (No. 429; CSE Technical Report). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Avenia-Tapper, B., & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment*, *20*(2), 95–111. https://doi.org/10.1080/10627197.2015.1028622

Bandalos, D. L. (2019). *Measurement theory and applications for the social sciences*. https://10.1080/15366367.2019.1610343

Banks, K., Jeddeeni, A., & Walker, C. M. (2016). Assessing the effect of language demand in bundles of math word problems. *International Journal of Testing*, *16*(4), 269–287. https://doi.org/10.1080/15305058.2015.1113972

Barrot, J. S. (2013). Revisiting the role of linguistic complexity in ESL reading comprehension. *3L: The Southeast Asian Journal of English Language Studies*, *19*(1), 5–18.

Brennan, R. L. (2001). *Generalizability theory*. Springer New York. https://doi.org/10.1007/978-1-4757-3456-0

Brooks, M. D. (2015). "It's Like a Script": Long-Term English Learners' Experiences with and Ideas about Academic Reading. *Research in the Teaching of English*, *49*(4), 383.

Buono, S., & Jang, E. E. (2021). The effect of linguistic factors on assessment of English language learners' mathematical ability: A differential item functioning analysis. *Educational Assessment*, *26*(2), 125–144. https://doi.org/10.1080/10627197.2020.1858783

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks* (CSE Report 642). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Callahan, R. M. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal*, *42*(2), 305-328. https://psycnet.apa.org/doi/10.3102/00028312042002305

Chen, J., Chen, C., & Shih, C. (2014). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement, 38*(1), 3-82. https://doi.org/10.1177/0146621613488643

Clinton, V., Basaraba, D. L., & Walkington, C. (2018). English learners and mathematical word problem solving: A systematic review. In *Second language acquisition: Methods, perspectives, & challenges* (pp. 171–208). Nova Science Publishers, Inc.

Cobb, T. *Web Vocabprofile* [accessed 26 October 2021 from http://lextutor.ca/vp/], an adaption of Heatley, Nation & Coxhead's (2002) *Range.*

Coxhead, A. (2000). A new academic word list. *TESOL quarterly, 34*(2), 213-238. https://doi.org/10.2307/3587951

Credé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: a critical review and development of reporting recommendations. *Journal of Organizational Behavior*, *36*(6), 845-872. https://psycnet.apa.org/doi/10.1002/job.2008

Darus, S., & Subramaniam, K. (2009). Error analysis of the written English essays of secondary school students in Malaysia: A case study. *European Journal of Social Sciences*, *8*(3), 483-495.

De Ayala, R. J. (2022). *The theory and practice of item response theory*. Guilford Publications.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models*. Springer New York. https://doi.org/10.1007/978-1-4757-3990-9

Dehghani, A. P., Bagheri, M. S., Sadighi, F., & Tayyebi, G. (2016). Investigating difficulty order of certain English grammatical features in an Iranian EFL setting. *International Journal of English Linguistics*, *6*(6), 209-220. http://dx.doi.org/10.5539/ijel.v6n6p209

Eckes, T. (2015). *Introduction to many-facet Rasch measurement* (Second). Peter Lang.

Faulkner-Bond, M., & Sireci, S. G. (2015). Validity issues in assessing linguistic minorities. *International Journal of Testing, 15*(2), 144-135. https://doi.org/10.1080/15305058.2014.974763

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *3*, 359-374.

García, O., Kleifgen, J. A., & Falchi, L. (2008). *From English language learners to emergent bilinguals* (No. 1; Equity Matters: Research Review). Teachers College, Columbia University.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, *28*, 24–34. https://doi.org/10.1016/j.learninstruc.2013.04.001

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. http://victoria.ac.nz/lals/staff/paul-nation.aspx

Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly*, *50*(1), 61–82. https://doi.org/10.1002/rrq.83

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, *10*(1), 128-141. https://doi.org/10.1207/S15328007SEM1001_6

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189-212). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-3990-9

Kachchaf, R., Noble, T., Rosebery, A., O'Connor, C., Warren, B., & Wang, Y. (2016). A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. *Bilingual Research Journal*, *39*(2), 152–166. https://doi.org/10.1080/15235882.2016.1169455

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differential of a state reading assessment: Item functioning, distractor functioning and omission frequency for disability categories. *Educational Measurement: Issues and Practice*, *28*(2), 28–40. https://psycnet.apa.org/doi/10.1111/j.1745-3992.2009.00145.x

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kieffer, M. J., & Parker, C. E. (2016). *Patterns of English learner student reclassification in New York City public schools* (REL 2017-200). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. http://ies.ed.gov/ncee/edlabs

Kim, W. G., & García, S. B. (2014). Long-term English language learners' perceptions of their language and academic learning experiences. *Remedial and Special Education*, *35*(5), 300-312. https://doi.org/10.1177/0741932514525047

Kline, R. B. (2023). Principles and practice of structural equation modeling. Guilford Press.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22-56. http://dx.doi.org/10.1177/0013164414529792

Lane, S., & Leventhal, B. (2015). Psychometric challenges in assessing English language learners and students with disabilities. *Review of Research in Education*, *39*(1), 165–214. https://doi.org/10.3102/0091732X14556073

Lee, M. K., & Randall, J. (2011). *Exploring language as a source of DIF in a math test for English language learners. NERA Conference Proceedings 2011, 20.* https://opencommons.uconn.edu/nera_2011/20

Liu, R., & Bradley, K. D. (2021). Differential item functioning among English language learners on a large-scale mathematics assessment. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.657335

Loughran, J. M. (2014). *Understanding differential item functioning for English language learners: The influence of linguistic complexity features* [Dissertation].

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *International Journal of Testing*, *8*(1), 14–33. https://doi.org/10.1080/15305050701808615

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*(2), 333–368. https://doi.org/10.17763/haer.78.2.70783570r1111t32

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*(3–4), 160–179. https://doi.org/10.1080/10627190903422906

Massachusetts Department of Elementary and Secondary Education (DESE). (2019a). MCAS 2019 released items biology, high school. https://www.doe.mass.edu/mcas/2019/release/hs-bio.pdf

Massachusetts Department of Elementary and Secondary Education (DESE). (2019b). MCAS 2019 released items mathematics, grade 10. https://www.doe.mass.edu/mcas/2019/release/gr10-math.pdf

Massachusetts Department of Elementary and Secondary Education (DESE). (2020a). *2019 Legacy MCAS Technical Report*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education (DESE). (2020b). *2019 Next-Generation MCAS and MCAS-Alt Technical Report*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education (DESE). (2022). *Guidance on English Leaner Education Services and Programming*. https://www.doe.mass.edu/ele/guidance/services-programming.docx

Menken, K., Kleyn, T., & Chae, N. (2012). Spotlight on "long-term English language learners": Characteristics and prior schooling experiences of an invisible population. *International Multilingual Research Journal*, *6*(2), 121-142. https://doi.org/10.1080/19313152.2012.665822

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement* (pp. 13-104). New York: American Council on Education and Macmillan.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds), *Explanatory item response models* (pp. 213-240). Springer New York. https://doi.org/10.1007/978-1-4757-3990-9

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109. https://doi.org/10.3102/1076998606298035

National Center for Education Statistics (NCES). (2021). English language learners in Public Schools. https://nces.ed.gov/programs/coe/indicator_cgf.asp.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3), 396-402. https://doi.org/10.3758/BF03200807

Oliveri, M. E. (2019). Considerations for designing accessible educational scenario-based assessments for multiple populations: A focus on linguistic complexity. *Frontiers in Education*, *4(*88). https://doi.org/10.3389/feduc.2019.00088

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286-300. https://doi.org/10.1080/08957347.2014.944305

Olsen, L. (2010). *Reparable harm: Fulfilling the unkept promise of educational opportunity for California's long term English learners.* Californians Together.

Olsen, L. (2014). Meeting the unique needs of long term English language learners. *National Education Association*.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education*, *16*(3), 223-243. https://psycnet.apa.org/doi/10.1207/S15324818AME1603_4

Pettersen, A., & Braeken, J. (2019). Mathematical competency demands of assessment items: A search for empirical evidence. *International Journal of Science and Mathematics Education*, *17*, 405-425. http://dx.doi.org/10.1007/s10763-017-9870-y

Plath, J., & Leiss, D. (2018). The impact of linguistic complexity on the solution of mathematical modelling tasks. *ZDM, 50*(1–2), 159–171. https://doi.org/10.1007/s11858-017-0897-x

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & Du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling.* Chapel Hill, NC: Scientific Software International, Inc.

Ravand, H. (2015). Item response theory using hierarchical generalized linear models. *Practical Assessment, Research, and Evaluation, 20*. https://doi.org/10.7275/s4n1-kn37

Riccardi, D., Lightfoot, J., Lam, M., Lyon, K., Roberson, N. D., & Lolliot, S. (2020). Investigating the effects of reducing linguistic complexity on EAL student comprehension in first-year undergraduate assessments. *Journal of English for Academic Purposes*, *43*, 100804. https://doi.org/10.1016/j.jeap.2019.100804

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. W. (2010). Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets. Final Report. NCEE 2009-4079. *National Center for Education Evaluation and Regional Assistance*.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323-338. https://doi.org/10.3200/JOER.99.6.323-338

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, *11*(2), 105–126. https://doi.org/10.1207/s15326977ea1102_2

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932. https://doi.org/10.1037/0003-066X.44.6.922

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184-199. https://doi.org/10.1177/0146621608321758

Shin, N. (2020). Stuck in the middle: Examination of long-term English learners. *International Multilingual Research Journal*, *14*(3), 181–205. https://doi.org/10.1080/19313152.2019.1681614

Shiu, J. L. (2011). EFL learners' perception of grammatical difficulty in relation to second language proficiency, performance, and knowledge. Dissertation. https://tspace.library.utoronto.ca/bitstream/1807/29869/1/Shiu_LiJu_201106_PhD_thesis.pdf

Sireci, S. G., Banda, E., & Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In S. Elliott, R. Kettler, P. Beddor, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and applications*, 231-246. Springer, Cham. https://doi.org/10.1007/978-3-319-71126-3_15

Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education*, *27*(4), 236–247. https://doi.org/10.1080/08957347.2014.944308

Solano-Flores, G., Barnett-Clarke, C., & Kachchaf, R. R. (2013). Semiotic structure and meaning making: The performance of English language learners on mathematics tests.

Educational Assessment, *18*(3), 147–161. https://doi.org/10.1080/10627197.2013.814515

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, *28*(2), 9-18. http://dx.doi.org/10.1111/j.1745-3992.2009.00143.x

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (2014). Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment*, *19*(4), 267–283. https://doi.org/10.1080/10627197.2014.964116

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53-75. https://doi.org/10.3102/10769986027001053

Taasoobshirazi, G., & Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods*, *11*(3), 31–39.

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1193–1208. https://doi.org/10.1044/1092-4388(2006/086)

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education*, *34*(15), 2343–2369. https://doi.org/10.1080/09500693.2012.705046

U.S. Department of Education, National Center for Education Statistics [NCES]. (2021). *The condition of education 2021* (2021-144), English Language Learners in Public Schools.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443–464. https://doi.org/10.3102/10769986030004443

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics.* Elsevier. https://doi.org/10.1016/S0169-7161(06)26004-8.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement*, *30*(1), 22-42. https://doi.org/10.1177/0146621605279867

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*(3–4), 139–159. https://doi.org/10.1080/10627190903425883

Young, J. W. (2008). Ensuring valid content tests for English language learners. *R&D Connections*, *8*.

Zieky, M. J. (2015). Developing fair tests. In S. Lane, M. S. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed.). Florence, KY: Routledge. https://doi.org/10.4324/9780203102961-11

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Appendix A

Coding Grammatical Features

Fill in all identification information at the top of the coding form. Be complete in the "Assessment Title" including grade, year, and subject. Upon completion of coding an item, confirm that the coding forms are properly marked with page numbers.

You may code all grammatical complexity features on one copy of the test. Additional copies may be used for clarity of markings as deemed necessary by the raters. In order to systematically and accurately identify and count the features as you progress through the passages and coding, it is important to notate each grammatical structure as it is encountered in the item.

For each item, indicate on the coding form the total number of times that a feature is used in the "Total" column. Count the number of times that feature includes construct relevant vocabulary (math vocabulary on the math test and biology vocabulary on the biology test) and indicate the count in the "CR" column.

1. Begin with **passive** and **complex verb** counts and proceed in this manner: as you read the item, **cross out** each non-complex/active verb thereby making the passive and complex verbs more apparent. Passive voice should be underlined and marked **PV**, and complex verb forms should be underlined and marked **CV**.
2. From verbs, move to coding **subordinate** and **relative clauses**, underlining and marking them **SC** and **RC** respectively. At this point, the text has been marked for passive voice, complex verbs, relative, and subordinate clauses.
3. Underline each **complex noun phrase** and mark as **NP**.
4. It is possible that you will discover additional grammatical complexities that originally went unnoticed as you progress through coding each feature. Be certain to go back to the appropriate text copy to mark any newly found complexities and update your code form.

Figure A1. Sample Grammatical Complexity Coding Form

Grammatical Complexity Code Form										
Rater:										
Subject (circle): Math Biology							**Year (circle):** 2018 2019			
Item #	**Passive (PV) Count**		**Complex Verb (CV) Count**		**Subordinate (SC) Count**		**Relative (RC) Count**		**Noun Phrase (NP) Count**	
	Total	**CR**	**Total**	**CR**	**Total**	**CR**	**Total**	**CR**	**Total**	**CR**
1										
2										
3										

The sections that follow detail how to count each grammatical feature.

Passive Voice/Verbs
In sentences written in passive voice, the subject receives the verb's action, as shown in Table A1.

Table A1. Passive and Active Voices and Simple and Complex Examples

Voice	Example	Note
Passive	The boy was bitten by the dog.	The boy is the subject and he is acted upon by being bitten. The subject is not doing the action.
Active	The dog bit the boy.	The dog is the subject and it acts by biting. The subject is doing the action.
Reduced passive verb	How did the Spaniards react when first introduced to chocolate?	…when they were first introduced…
Reduced passive verb – part of reduced relative clause	The birds infected with West Nile Virus… The man arrested last night…	Code as RC only, not as a passive verb
Passive verb in a relative clause	The fruit, which will eventually be converted into chocolate…	Not reduced, count as both PV and RC

Examples of Passive Voice
- The chocolate gave them the strength to carry on until more food rations could be obtained.
- His wound was treated at the hospital.
- Used by small shops
- Was/were paid
- Is being read
- Will be published
- Was/were sold
- Had/has been computed
- Could be seen

Sample Coding

PV

Spanish monks, who had been consigned to process the cocoa beans, finally let the secret out.

For each item indicate on the coding form the total number of times that the passive voice is used in the "Total" column. Count the number of times that passive voice phrases include construct relevant vocabulary (math vocabulary on the math test and biology vocabulary on the biology test) and indicate the count in the "CR" column.

Complex Verbs

Complex verbs are multi-part with a base or main verb and several auxiliaries. Table A2 lists complex verbs and Table A3 shows multi-part verbs that are not counted as complex verbs.

Table A2. Complex Verb Forms.

Type	**Structure**	**Example**
present perfect continuous	have/has + been + present participle	has been waiting
past perfect continuous	had been + present participle	had been waiting
future continuous	will be + present participle	will be waiting
future continuous	am/is/are + going to be + present participle	are going to be waiting
future perfect continuous	will have been + present participle	will have been waiting
future perfect continuous	am/is/are + going to have been + present participle	are going to have been waiting
used to	used to + verb	used to go
present/past participle	have/had + participle + infinitive	have/had wanted to go was/were hoping to go
modals	modal + verb	can/could work, might run, should always go, ought to help, would help
subjunctive	if + subject + verb	if I were a rich person, whether it be true or false
future in the past	was/were + going to + verb	were going to go

Table A3. Not Complex Verb Forms.

Type	Structure	Example
simple present	verb, verb + s/es	wait, waits
present continuous	am/is/are + present participate	is dancing, are hurrying
simple past	verb + ed, or irregular verbs	waited, ran
simple past with "do"	did + verb	did take, did you take?
Past continuous	was/were + present participle	was dancing, were hurrying
present perfect	has/have + past participle	has become, have seen
past perfect	had + past participle	had studied
simple present/past	simple present/past verb + infinitive/participle	want/wanted to see, begin working
simple future	will + verb	will wait

Sample Coding

CV

But, only 3 to 10 percent <u>will go on to mature</u> into full fruit.

CV

Ultimately, someone decided the drink <u>would taste</u> better if served hot.

For each item indicate on the coding form the total number of times that a complex verb is used in the "Total" column. Count the number of times that complex verbs include construct relevant vocabulary (math vocabulary on the math test and biology vocabulary on the biology test) and indicate the count in the "CR" column. <u>Do not count passive voice verbs as complex verbs.</u>

Relative Clauses
A relative clause is one type of subordinate clause that modifies a noun or pronoun by identifying or classifying it. It is also called an adjective clause and nearly always follows the word modified. It is introduced by a relative pronoun. Examples of relative clause types are shown in Table A4. Relative pronouns and adverbs are shown in Table A5.

Relative clauses generally meet four criteria –
1) They contain a subject and a verb,
2) They begin with a relative pronoun,
3) They answer the questions: What kind? How many? Which one?
4) They do not form a complete sentence.

Table A4. Relative Clause Patterns and Sample Coding.

Relative clause type	**Example**	**Note**
Relative pronoun + subject + verb	Cacao trees get their start in a nursery bed <u>where</u> (relative pronoun) <u>seeds</u> (noun) from high-yielding <u>trees are planted</u> (verb-passive) in fiber baskets or plastic bags.	The relative clause modifies the noun "nursery bed" by identifying which nursery bed. Count as RC and PV.
Relative pronoun as subject + verb	Spain wisely proceeded to plant cocoa in its overseas colonies, <u>which</u> (relative pronoun as subject) <u>gave</u> (verb) <u>birth</u> to a very profitable business.	The relative clause modifies the noun "colonies" by identifying which colony.
Reduced relative clause (missing relative pronoun + adverbial verb)	From then on, drinking chocolate had more of the smooth consistency and the pleasing flavor <u>it</u> (subject) <u>has</u> (verb) <u>today</u>.	"That" is omitted: "…that it has today."
Relative clause with passive verb	The fruit, <u>which will eventually be converted into chocolate</u>…	Not reduced, count as both RC and PV.

Table A5. Relative Pronouns.

Relative Pronouns		
that	whoever	whomever
which	whomever	where
whichever	whose	where
who	whosever	why

Examples of Relative Clauses

The money <u>which Francine did not accept</u> was given as a gift.
(which = relative pronoun, Francine = subject, did accept = verb)

George went to the flea market <u>where he found the baseball card in good condition.</u>
(which = relative pronoun, he = subject, found = verb)

There was her necklace <u>that dangled from the edge of the cabinet.</u>
(that = relative pronoun as a subject, dangled = verb)

The man <u>I lent my car to</u> last night is my neighbor.
(reduced relative clause – null, pronoun = “who” is dropped/omitted, I = subject, lent = verb)

He devised a way of adding milk to the chocolate, creating the product <u>we enjoy today known as milk chocolate.</u>
(Two null relative clauses: “that” is dropped/omitted, we = subject, enjoy = verb, and “that is” is dropped/omitted, known = verb – “that we enjoy today that is known as milk chocolate.”)

For each item indicate on the coding form the total number of relative clauses in the “Total” column. Count the number of times relative clauses include construct relevant vocabulary (math vocabulary on the math test and biology vocabulary on the biology test) and indicate the count in the “CR” column.

Subordinate/Dependent Clauses
Other subordinate clauses that are NOT relative clauses. Other subordinate clauses function within the sentence as a noun or an adverb. Table A6 shows subordinate conjunctions.

Subordinate clauses usually meet four criteria:
1) They contain a subject and a verb.
2) They begin with a subordinate conjunction.
3) They do not form a complete sentence.
4) They act as a noun or adverb.

Table A6. Subordinate Conjunctions

Subordinate Conjunctions		
after	once	until
although	provided that	when
as	rather than	whenever
because	since	where
before	so that	whereas
even if	than	wherever
even though	that	whether
if	though	while
in order that	unless	why

Examples of Subordinate Clauses
<u>After he threw the ball</u>, the outfielder yelled to the first baseman.
The subordinate clause functions as an adverb to answer the question "when."

Some say <u>it originated in the Amazon basin of Brazil</u>, <u>while still others contend</u> <u>that it is native to Central America.</u> (three subordinate clauses beginning with the conjunctions "that" understood as "that it originated in the Amazon Basin of Brazil," "while," and "that.")

<u>To make the concoction more agreeable to Europeans</u>, Cortez and…
("In order" is understood: "In order to make the concoction…")

We know <u>it does not matter.</u>

Each year, <u>as the article says,</u> draws a crowd.

For each item indicate on the coding form the total number of subordinate clauses that are not relative clauses in the "Total" column. Count the number of times subordinate clauses include construct relevant vocabulary (math vocabulary on the math test and biology vocabulary on the biology test) and indicate the count in the "CR" column.

Complex Noun Phrase
The main structure in the phrase is the noun, but the addition of determiners, adjectives/modifiers, and prepositional phrases adds complexity. Table A7 gives examples.

Table A7. Noun Phrases.

Y/N	Structure	Example
Yes	determiner + three or more modifiers + noun	The old straggly red chickens
Yes	determiner + modifier + noun + prepositional phrase	The red chickens in the coup
Yes	three or more modifiers + noun	Tiny waxy pink blossoms…
Yes	modifier + noun + prepositional phrase	The hot valleys of Southern California…
Yes	noun + two prepositional phrases	The valleys of Southern California in the summer…
Yes	noun + noun	Electron microscope, furniture replacement, New World offerings
No	noun	Chickens
No	determiner + noun	The chickens
No	determiner + modifier + noun	The red chickens
No	modifier + noun	Red chickens

Count each word separately in hyphenated modifiers. For example, "rich, well-drained soil" is a complex noun phrase because it consists of a noun (soil) and three modifiers (rich, well, and drained).

A noun phrase within a noun phrase counts as only one complex noun phrase. For example: The 19^{th} century marked two more revolutionary *developments in the history of chocolate.*
The underlined complex noun phrase, "two more revolutionary developments" (3 modifiers + noun) is also part of the italics noun phrase "developments in the history" (noun + prepositional phrase) which includes another noun phrase, "history of chocolate" (noun + prepositional phrase). The entire phrase from "two" through "chocolate" is counted as only one complex noun phrase.
A noun phrase that is identically repeated within the same paragraph is counted only once.

Proper noun + noun: count first time only in passage. Example: the game Rocket Ball.
Common noun + common noun. Count three times max in the passage. Example: cacao tree.

Please err on not over-counting noun + noun. Skip proper nouns such as someone's name or U.S. Government.

Examples and Sample Coding

The story of chocolate, as far back as we know it, begins with the discovery of America.

The hand methods of manufacture used by small shops gave way in time to the mass production of chocolate.

A newly planted cacao seedling is often sheltered by a different type of tree.

Table A8 lists frequently used prepositions to aid in the identification of noun phrases that include a prepositional phrase.

Table A8.

Examples of Prepositions				
about	below	excepting	off	toward
above	beneath	for	on	under
across	beside(s)	from	onto	underneath
after	between	in	out	until
against	beyond	in front of	outside	up
along	but	inside	over	upon
among	by	in spite of	past	up to
around	concerning	instead of	regarding	with
at	despite	into	since	within
because of	down	like	through	without
before	during	near	throughout	with regard to
behind	except	of	to	with respect to

For each item indicate on the coding form the total number of complex noun phrases in the "Total" column. Count the number of times noun phrases include construct relevant vocabulary (math vocabulary on the math test and biology vocabulary on the biology test) and indicate the count in the "CR" column

Appendix B

MCAS Biology Construct Relevant Words

10% rule of energy transfer
abiotic
abiotic resource
activation energy
active transport
active transport potential
adaptation
aerobic cellular respiration
alleles
alveoli
amino acid
anatomical
anatomy
artery
asexual
asexual reproduction
atmosphere
ATP
average
bacterium
behavioral
biochemical
biochemical reaction
biodiversity
biological
biological communities
biomass
biosphere
biotic
birth
blood
blood cell
blood clotting
body
body function
bond
bones
brain
capillary
captive breeding program
carbohydrate
carbon
carbon dioxide
carnivore
cartilage
catalyst
cell
cell biology
cell cycle
cell growth
cell membrane
cell part
cell wall
cell waste
cellular respiration
centriole
chemical energy
chemical reaction
chemistry of life
chloroplast
chromosome
cilium
circulatory
circulatory system
class
climate
climate change
codominant
combustion
commensalism
comparative anatomy
competition
complementary base
complementary nucleotide pair
compound
concentration gradient
conservation
consumer
crossing over
cytoplasm
cytoskeleton
Darwin's theory of evolution
death
decomposer
decomposition
deoxyribonic nucleic acid
diaphragm
diffusion
digestive
digestive system
dihybrid cross
diploid
diploid zygote
disaccharide
disease
DNA
DNA replication
DNA sequence
dominant
dominant-recessive
double helix
double-stranded
ecology
ecosystem
ecotourism
element
electrochemical signals
emigration
endoplasmic reticulum
energy
energy conservation
energy pyramid
energy transfer
enzyme
esophagus
evidence
evolution
excretory
excretory function
express
expressed trait
extinction
facilitated diffusion
family

fats
fatty acid
feedback mechanism
fertilization
flagellum
food web
fossil
fossil record
fungi
fungus
gamete
gene
gene expression
gene flow
genetic code
genetic diversity
genetic drift
genetic information
genetic inheritance
genetic material
genetic trait
genetic variation
genetics
genome
genotype
genus
geographic isolation
geosphere
Golgi apparatus
habitat
habitat fragmentation
habitat restoration
haploid cell
heart
hemoglobin
herbivore
heritable
hierarchical taxonomic system
homeostasis
homologous
homology
hormone
human activity
hydrocarbons
hydrogen
hydrosphere
immigration
incomplete dominance
independent assortment
inherit
inheritance
inheritance pattern
inorganic compound
invasive species
ions
kidney
kingdom
large intestine
larynx
light energy
lipid
liver
lungs
lysosome
macromolecule
mediate
meiosis
Mendel
Mendelian inheritance
metabolism
microorganism
mitochondrion
mitosis
molecular
molecular biology
molecular structure
molecule
monohybrid cross
monomer
monosaccharide
morphological
motor neuron
mouth
multiple alleles
muscle
mutation
mutualism
natural causes
natural disaster
natural selection
negative feedback
nerve
nervous system
neuron
nitrogen
nitrogenous waste
non-native species
nose
nuclear envelope
nuclear membrane
nucleic acid
nucleolus
nucleotide
nucleus
nutrient uptake
nutrients
offspring
order
organelle
organic matter
organic molecule
organism
osmosis
overharvesting
oxygen
pancreas
parasitism
passive transport
pedigree chart
pH
pharynx
phenotype
phenotypic change
phosphate
phosphate backbone
phosphorus
photosynthesis
phylum
physiological feedback loop
physiology
plasma membrane
platelet
pollution
pollution mitigation
polygenic
polysaccharide

population
positive feedback
predation
primary function
probability
producer
product
protein
protist
pseudopod
Punnett Square
pyramid (energy)
reactant
reaction
receptor
recessive
rectum
red blood cell
replication
reproduction
respiration
respiratory system
ribosome
RNA
segration
selective barrier
sensory neuron
sequences (amino acid)
sequences (genetic)
sequences (nucleic acid)
sex-linked
sexual reproduction
sexually produced
offspring
skin
small intestine
speciation
species
species diversity
spinal cord
stomach
structural protein
structure
sugars
sulfur
symbioses
synthesis (protein, glycose)
taxonomy
temperature
trachea
trait
transcription
translation
transmission
trend
triglyceride
trophic level
vacuole
vein
vertebrates
vestigial
villi
virus
waste (dead organic
material)
water
web (food)
zygote

Appendix C

MCAS Mathematics Construct Relevant Words

absolute value
acute angle
add
algebra
amplitude
angle (geometry)
appreciation (value)
arc
area (of surface/shape)
arithmetic sequence
associative property
average
base (log)
base angle
bisector
box plot
calculate
categorical data
central angle
chord (circle)
circle (geometry)
circumference
circumscribed angle
circumscribed circle
coefficient
commutative property
compass
complement
complementary angles
complex number
complex solution
congruence
conditional frequency
conditional probability
cone
constant term
constraint
coordinate axis
coordinate pair
coordinate plane
correlation
correlation coefficient
corresponding angle
corresponding pair
cosine
counterclockwise
cross section
cube root
curve
cylinder
data
data distribution
degree
density
depreciation (value)
diagonal (parallelogram)
difference of squares
dilation
directrix (parabola)
distance formula
distributive property
division
domain
dot plot
element
end behavior
endpoint
equation
equidistant
equilateral triangle
equivalent
error
experiment
explicit expression
exponent
exponential
exponential function
expression
exterior angle
factor
fitted function
focus (parabola)
formula
frequency
frequency table
function
geometric
geometric sequence
geometry
graph
half-plane
histogram
horizontal stretch
independent (probability)
inequality
inference
input
input-output pair
inscribe
inscribed angle
inscribed circle
inscription
integer
integer exponent
intercept
interior angle
interpret
interquartile range
intersect
intersection
interval
inverse
inverse function
irrational
irrational number
isosceles triangle
joint frequency
label
length
line segment
linear
linear function
logarithm
logarithmic
long division
margin of error

marginal frequency
maxima
maximum
mean (average)
measurement
median
midline
midpoint
minima
minimum
multiply
multi-step
negative
nonzero
normal distribution
notation
number
number line
numerical relationship
observational study
outcome
outlier
output
pairs of equations
parabola
parallel line
parallelogram
parameter
perimeter
periodicity
perpendicular line
plane (coordinate)
plot
point
polygon
polynomial
polynomial identities
positive
product
property
proportionally
pyramid
Pythagorean
quadrant
quadratic
quadratic formula
quadratically
quadrilateral
quantile
quantity
radian
radical
radius
randomization
randomized experiment
range
rate of change
rate per unit
ratio
rational
rational
rational exponent
rational expression
rational number
real number
rectangle
recursive process
reflection
relationship
relative frequency
relative maximum
relative minimum
remainder
remainder theorem
residual
right triangle
rigid motion
root function
rotation
rounding
sample
scale
scale factor
scatter plot
segment
sequence
side ratio
similar (angle)
sine
slope (line)
solution
sphere
square root
standard deviation
statistic
step function
straightedge ruler
subtract
sum
survey
symmetry
systems of equations
table
tangent (circle)
term
theorem
three-dimensional
transformation
translation
transversal
trapezoid
treatment (experiment)
triangle
trigonometric
trigonometric ratio
two-dimensional
union
unit
unit circle
variable
vertical angle
volume (of object)
width
x-coordinate
zeros

Appendix D

Features of MCAS Assessments

This appendix contains tables for the features of the MCAS assessements used in the present study. Tables D1 and D2 present the item score descriptive statistics for the mathematics and biology assessments, respectively. Tables D3 and D4 present the item type, points possible and reporting categories for the mathematics and biology assessments, respectively. Tables D5 and D6 present the comparison group by item score correlations for the mathematics and biology assessments, respectively. Tables D7 and D8 present the lexical complexity, complex noun phrases, and relative clauses factor scores for each item for the mathematics and biology assessments, respectively.

Table D1.

Item Score Descriptive Statistics – Mathematics Assessment

Item	Mean	Standard Deviation	Item	Mean	Standard Deviation
m01	0.886	0.318	m22	0.641	0.480
m02	0.598	0.490	m23	0.839	0.368
m03	0.658	0.474	m24	0.499	0.500
m04	0.586	0.493	m25	0.835	0.371
m05	0.569	0.495	m26	0.662	0.473
m06	0.861	0.346	m27	0.725	0.447
m07	0.371	0.483	m28	0.533	0.499
m08	0.571	0.495	m29	0.601	0.490
m09	1.646	1.166	m30	2.042	1.168
m10	0.614	0.487	m31	0.838	0.369
m11	0.593	0.491	m32	0.483	0.500
m12	1.080	0.804	m33	1.666	0.578
m13	0.640	0.480	m34	0.583	0.493
m14	2.054	1.303	m35	1.989	1.613
m15	0.485	0.500	m36	0.727	0.446
m16	1.372	0.763	m37	0.930	0.757
m17	0.585	0.493	m38	0.733	0.442
m18	0.494	0.500	m39	0.564	0.496
m19	0.964	0.802	m40	1.344	0.742
m20	0.497	0.500	m41	0.582	0.493
m21	0.622	0.485	m42	0.743	0.437

Table D2.

Item Score Descriptive Statistics – Biology Assessment

Item	Mean	Standard Deviation	Item	Mean	Standard Deviation
b1	0.747	0.435	b24	0.726	0.446
b2	0.683	0.465	b25	0.757	0.429
b3	0.621	0.485	b26	0.595	0.491
b4	0.669	0.471	b27	0.758	0.428
b5	0.593	0.491	b28	0.786	0.410
b6	0.707	0.455	b29	0.777	0.416
b7	0.685	0.464	b30	0.657	0.475
b8	0.852	0.355	b31	0.706	0.456
b9	0.680	0.467	b32	1.345	1.155
b10	0.500	0.500	b33	0.438	0.496
b11	0.729	0.444	b34	0.755	0.430
b12	1.465	1.044	b35	0.572	0.495
b13	0.735	0.441	b36	0.716	0.451
b14	0.738	0.440	b37	0.745	0.436
b15	0.681	0.466	b38	0.674	0.469
b16	0.649	0.477	b39	0.709	0.454
b17	0.421	0.494	b40	0.737	0.440
b18	0.663	0.473	b41	0.637	0.481
b19	0.624	0.484	b42	0.772	0.420
b20	0.589	0.492	b43	0.625	0.484
b21	0.560	0.496	b44	1.340	0.972
b22	0.512	0.500	b45	1.995	1.337
b23	1.336	1.008			

Table D3.

Item Type, Points Possible, and Reporting Category – Mathematics Assessment

Item	Item Type	Points Possible	Reporting Category	Item	Item Type	Points Possible	Reporting Category
m01	SR	1	A	m22	SR	1	G
m02	SR	1	A	m23	SR	1	S
m03	SR	1	G	m24	SR	1	A
m04	SR	1	G	m25	SR	1	G
m05	SA	1	A	m26	SR	1	G
m06	SR	1	A	m27	SR	1	G
m07	SR	1	A	m28	SR	1	A
m08	SR	1	G	m29	SA	1	A
m09	CR	4	N	m30	CR	4	G
m10	SR	1	S	m31	SR	1	A
m11	SR	1	G	m32	SR	1	G
m12	SR	2	S	m33	SA	2	N
m13	SR	1	A	m34	SR	1	G
m14	CR	4	S	m35	CR	4	A
m15	SR	1	N	m36	SR	1	G
m16	SR	2	G	m37	SR	2	A
m17	SR	1	A	m38	SR	1	A
m18	SR	1	G	m39	SR	1	G
m19	SR	2	N	m40	SR	2	A
m20	SR	1	A	m41	SR	1	G
m21	SR	1	G	m42	SR	1	S

Note: SR = selected response, SA = short answer, CR = constructed response, A = Algebra and Functions, G = Geometry, N = Number and Quantity, S = Statistics and Probability.

Table D4.

Item Type, Points Possible, and Reporting Category – Biology Assessment

Item	Item Type	Points Possible	Reporting Category	Item	Item Type	Points Possible	Reporting Category
b01	MC	1	Gen	b24	MC	1	Evo
b02	MC	1	Cell	b25	MC	1	Eco
b03	MC	1	Eco	b26	MC	1	Cell
b04	MC	1	Evo	b27	MC	1	Cell
b05	MC	1	Eco	b28	MC	1	Eco
b06	MC	1	Eco	b29	MC	1	Cell
b07	MC	1	AP	b30	MC	1	Gen
b08	MC	1	Eco	b31	MC	1	AP
b09	MC	1	Evo	b32	CR	4	Cell
b10	MC	1	Evo	b33	MC	1	Gen
b11	MC	1	Eco	b34	MC	1	AP
b12	CR	4	Eco	b35	MC	1	Cell
b13	MC	1	Gen	b36	MC	1	Evo
b14	MC	1	AP	b37	MC	1	Cell
b15	MC	1	AP	b38	MC	1	Cell
b16	MC	1	Evo	b39	MC	1	Gen
b17	MC	1	Gen	b40	MC	1	Eco
b18	MC	1	Cell	b41	MC	1	Gen
b19	MC	1	Cell	b42	MC	1	Evo
b20	MC	1	Cell	b43	MC	1	Cell
b21	MC	1	Evo	b44	CR	4	Evo
b22	MC	1	Gen	b45	CR	4	Gen
b23	CR	4	AP				

Note: MC = multiple choice, CR = constructed response, AP = Anatomy and Physiology, Cell = Biochemistry and Cell Biology, Eco = Ecology, Evo = Evolution and Biodiversity, Gen = Genetics

Table D5.

Comparison Group by Item Score Correlations – Mathematics Assessment

Item	EPvEB	EPvSTEB	EPvLTEB	STEB vLTEB	EPvSPA	EPvOTH	OTH vSPA
m01	-0.131*	-0.103*	-0.093*	-0.051*	-0.127*	-0.061*	-0.104*
m02	-0.182*	-0.151*	-0.111*	-0.029	-0.169*	-0.090*	-0.196*
m03	-0.157*	-0.127*	-0.101*	-0.047*	-0.140*	-0.085*	-0.104*
m04	-0.171*	-0.143*	-0.101*	-0.004	-0.152*	-0.091*	-0.140*
m05	-0.191*	-0.158*	-0.116*	-0.029	-0.166*	-0.107*	-0.137*
m06	-0.147*	-0.122*	-0.093*	-0.019	-0.144*	-0.066*	-0.131*
m07	-0.124*	-0.099*	-0.080*	-0.063*	-0.105*	-0.073*	-0.065*
m08	-0.150*	-0.119*	-0.099*	-0.066*	-0.132*	-0.082*	-0.103*
m09	-0.207*	-0.171*	-0.127*	-0.035	-0.189*	-0.106*	-0.237*
m10	-0.130*	-0.109*	-0.077*	-0.005	-0.115*	-0.071*	-0.082*
m11	-0.122*	-0.100*	-0.076*	-0.028	-0.106*	-0.069*	-0.064*
m12	-0.204*	-0.170*	-0.123*	-0.027	-0.167*	-0.128*	-0.050*
m13	-0.172*	-0.147*	-0.099*	0.010	-0.152*	-0.094*	-0.111*
m14	-0.264*	-0.223*	-0.156*	0.001	-0.227*	-0.153*	-0.201*
m15	-0.114*	-0.095*	-0.068*	-0.012	-0.095*	-0.069*	-0.037
m16	-0.194*	-0.156*	-0.127*	-0.068*	-0.178*	-0.099*	-0.164*
m17	-0.142*	-0.110*	-0.099*	-0.087*	-0.133*	-0.069*	-0.149*
m18	-0.180*	-0.152*	-0.105*	0.001	-0.153*	-0.106*	-0.111*
m19	-0.160*	-0.126*	-0.106*	-0.089*	-0.134*	-0.095*	-0.071*
m20	-0.078*	-0.057*	-0.058*	-0.068*	-0.069*	-0.041*	-0.057*
m21	-0.129*	-0.103*	-0.085*	-0.048*	-0.121*	-0.062*	-0.127*
m22	-0.099*	-0.071*	-0.077*	-0.094*	-0.080*	-0.063*	-0.011
m23	-0.175*	-0.149*	-0.105*	-0.003	-0.162*	-0.090*	-0.110*
m24	-0.012*	-0.008*	-0.011*	-0.019	-0.013*	-0.004*	-0.020
m25	-0.199*	-0.181*	-0.101*	0.068*	-0.168*	-0.123*	-0.043*
m26	-0.121*	-0.097*	-0.081*	-0.047*	-0.108*	-0.066*	-0.074*
m27	-0.185*	-0.159*	-0.107*	0.009	-0.170*	-0.096*	-0.139*
m28	-0.085*	-0.072*	-0.049*	0.003	-0.079*	-0.040*	-0.087*
m29	-0.213*	-0.176*	-0.132*	-0.048*	-0.180*	-0.128*	-0.095*
m30	-0.273*	-0.231*	-0.163*	0.003	-0.231*	-0.166*	-0.120*
m31	-0.288*	-0.251*	-0.169*	0.021	-0.258*	-0.165*	-0.127*
m32	-0.130*	-0.104*	-0.084*	-0.052*	-0.118*	-0.066*	-0.126*
m33	-0.272*	-0.230*	-0.173*	-0.026	-0.249*	-0.149*	-0.150*
m34	-0.097*	-0.078*	-0.063*	-0.033	-0.089*	-0.048*	-0.085*
m35	-0.234*	-0.194*	-0.143*	-0.055*	-0.201*	-0.136*	-0.207*

Item	EPvEB	EPvSTEB	EPvLTEB	STEB vLTEB	EPvSPA	EPvOTH	OTH vSPA
m36	-0.182*	-0.152*	-0.111*	-0.019	-0.163*	-0.100*	-0.108*
m37	-0.172*	-0.141*	-0.106*	-0.042*	-0.145*	-0.101*	-0.085*
m38	-0.129*	-0.106*	-0.082*	-0.029	-0.126*	-0.057*	-0.138*
m39	-0.126*	-0.103*	-0.078*	-0.027	-0.105*	-0.076*	-0.041
m40	-0.205*	-0.171*	-0.124*	-0.023	-0.179*	-0.116*	-0.119*
m41	-0.157*	-0.125*	-0.104*	-0.069*	-0.136*	-0.089*	-0.091*
m42	-0.181*	-0.157*	-0.103*	0.018	-0.152*	-0.112*	-0.046*

Note: * = p < .01.

Table D6.

Comparison Group by Item Score Correlations – Biology Assessment

Item	EPvEB	EPvSTEB	EPvLTEB	STEB vLTEB	EPvSPA	EPvOTH	OTH vSPA
b1	-0.234*	-0.207*	-0.141*	-0.015	-0.213*	-0.137*	-0.073*
b2	-0.189*	-0.164*	-0.115*	-0.023	-0.173*	-0.106*	-0.074*
b3	-0.140*	-0.125*	-0.079*	0.000	-0.124*	-0.083*	-0.036
b4	-0.176*	-0.162*	-0.091*	0.030	-0.153*	-0.110*	-0.028
b5	-0.141*	-0.128*	-0.075*	0.016	-0.126*	-0.082*	-0.044
b6	-0.138*	-0.123*	-0.077*	0.005	-0.131*	-0.070*	-0.074*
b7	-0.099*	-0.076*	-0.077*	-0.068*	-0.092*	-0.053*	-0.046
b8	-0.243*	-0.223*	-0.138*	0.017	-0.238*	-0.125*	-0.116*
b9	-0.214*	-0.187*	-0.130*	-0.024	-0.185*	-0.135*	-0.031
b10	-0.155*	-0.139*	-0.083*	0.014	-0.141*	-0.085*	-0.073*
b11	-0.261*	-0.229*	-0.161*	-0.028	-0.247*	-0.140*	-0.125*
b12	-0.338*	-0.306*	-0.182*	0.069*	-0.296*	-0.205*	-0.124*
b13	-0.234*	-0.210*	-0.135*	0.000	-0.225*	-0.120*	-0.129*
b14	-0.169*	-0.155*	-0.091*	0.019	-0.171*	-0.074*	-0.130*
b15	-0.189*	-0.171*	-0.104*	0.013	-0.175*	-0.103*	-0.083
b16	-0.190*	-0.162*	-0.121*	-0.042	-0.175*	-0.105*	-0.081
b17	-0.093*	-0.082*	-0.052*	-0.001	-0.077*	-0.060*	-0.004
b18	-0.176*	-0.156*	-0.101*	-0.003	-0.164*	-0.093*	-0.086
b19	-0.122*	-0.088*	-0.105*	-0.119*	-0.113*	-0.064*	-0.059
b20	-0.121*	-0.102*	-0.078*	-0.035	-0.106*	-0.072*	-0.030
b21	-0.093*	-0.080*	-0.057*	-0.016	-0.089*	-0.046*	-0.060*
b22	-0.111*	-0.101*	-0.056*	0.020	-0.105*	-0.055*	-0.069*
b23	-0.281*	-0.244*	-0.165*	-0.019	-0.255*	-0.155*	-0.185*
b24	-0.203*	-0.176*	-0.127*	-0.030	-0.190*	-0.109*	-0.090
b25	-0.295*	-0.279*	-0.146*	0.076	-0.278*	-0.162*	-0.130*
b26	-0.152*	-0.123*	-0.107*	-0.073*	-0.145*	-0.076*	-0.094
b27	-0.171*	-0.152*	-0.101*	-0.007	-0.165*	-0.086*	-0.093
b28	-0.346*	-0.329*	-0.177*	0.084	-0.317*	-0.210*	-0.097
b29	-0.194*	-0.157*	-0.142*	-0.084	-0.191*	-0.092*	-0.121*
b30	-0.103*	-0.084*	-0.072*	-0.046	-0.107*	-0.039*	-0.104*
b31	-0.175*	-0.151*	-0.110*	-0.031	-0.175*	-0.079*	-0.131*
b32	-0.245*	-0.206*	-0.152*	-0.071*	-0.223*	-0.131*	-0.184*
b33	-0.101*	-0.096*	-0.044*	0.046	-0.091*	-0.057*	-0.039
b34	-0.225*	-0.209*	-0.116*	0.041	-0.193*	-0.148*	-0.016
b35	-0.129*	-0.111*	-0.079*	-0.021	-0.123*	-0.064*	-0.082

b36	-0.268*	-0.243*	-0.148*	0.021	-0.244*	-0.155*	-0.093
b37	-0.213*	-0.177*	-0.149*	-0.076	-0.194*	-0.124*	-0.066*
b38	-0.187*	-0.166*	-0.106*	0.000	-0.177*	-0.096*	-0.105*
b39	-0.106*	-0.088*	-0.071*	-0.036	-0.101*	-0.053*	-0.060*
b40	-0.225*	-0.198*	-0.136*	-0.018	-0.206*	-0.129*	-0.077
b41	-0.131*	-0.118*	-0.071*	0.009	-0.117*	-0.076*	-0.041
b42	-0.284*	-0.257*	-0.163*	0.010	-0.270*	-0.152*	-0.134*
b43	-0.108*	-0.083*	-0.084*	-0.078	-0.103*	-0.053*	-0.067*
b44	-0.296*	-0.266*	-0.161*	0.051	-0.265*	-0.169*	-0.159*
b45	-0.250*	-0.212*	-0.159*	-0.064*	-0.231*	-0.135*	-0.150*

Note: * = $p < .01$.

Table D7.

Linguistic Feature Factor Scores by Item – Mathematics Assessment

Item	Lexical Complexity	Complex Noun Phrases	Relative Clauses	Item	Lexical Complexity	Complex Noun Phrases	Relative Clauses
m01	-0.920	-0.622	-0.489	m22	-0.508	-0.216	-0.489
m02	-0.714	-0.622	-0.489	m23	0.876	0.067	-0.489
m03	-0.546	-0.369	-0.489	m24	-0.920	-0.664	-0.489
m04	0.390	-0.636	-0.489	m25	-0.658	-0.608	-0.489
m05	-0.490	0.164	-0.489	m26	-0.752	-0.664	-0.489
m06	-0.920	-0.622	-0.489	m27	-0.639	-0.469	-0.489
m07	-0.883	-0.664	-0.489	m28	1.063	-0.608	-0.489
m08	-0.696	-0.664	-0.489	m29	-0.003	0.640	-0.489
m09	1.849	-0.664	-0.489	m30	1.063	-0.523	2.933
m10	1.287	0.514	0.652	m31	1.306	0.246	-0.489
m11	-0.677	-0.664	-0.489	m32	-0.696	-0.608	-0.489
m12	1.138	-0.331	2.933	m33	0.839	1.530	-0.489
m13	-0.957	-0.636	-0.489	m34	-0.415	-0.538	-0.489
m14	1.998	1.216	2.933	m35	3.626	4.151	2.933
m15	-0.228	0.246	0.652	m36	0.184	0.866	0.652
m16	-0.116	-0.031	-0.489	m37	1.456	-0.157	-0.489
m17	-0.920	-0.622	-0.489	m38	0.520	2.872	-0.489
m18	-0.247	-0.636	-0.489	m39	-0.303	-0.664	-0.489
m19	1.194	-0.031	-0.489	m40	1.606	3.289	0.652
m20	-0.883	-0.594	-0.489	m41	-0.621	-0.594	-0.489
m21	-0.658	-0.664	-0.489	m42	0.708	-0.367	-0.489

Table D8.

Linguistic Feature Factor Scores by Item – Biology Assessment

Item	Lexical Complexity	Complex Noun Phrases	Relative Clauses	Item	Lexical Complexity	Complex Noun Phrases	Relative Clauses
b01	2.105	0.365	2.232	b24	0.505	0.985	-0.342
b02	-1.735	-0.739	-0.300	b25	-0.775	-0.363	-0.380
b03	0.078	0.323	-0.380	b26	-1.308	-0.709	-0.380
b04	0.292	-0.326	-0.380	b27	1.358	-0.357	-0.361
b05	0.078	-0.484	-0.380	b28	-0.135	1.480	3.026
b06	-0.988	0.381	-0.380	b29	-1.308	-0.333	-0.380
b07	-0.882	-0.426	-0.380	b30	-0.028	-0.333	-0.380
b08	0.398	-0.611	-0.371	b31	-0.348	-0.611	-0.380
b09	1.465	-0.087	0.490	b32	0.292	0.288	-0.380
b10	1.465	-0.552	0.499	b33	-1.095	-0.709	-0.380
b11	-1.415	-0.332	-0.380	b34	-1.095	-0.611	-0.361
b12	1.465	0.660	-0.380	b35	0.611	-0.395	-0.380
b13	-0.775	-0.611	-0.380	b36	0.398	1.086	2.241
b14	0.185	0.009	0.533	b37	-0.455	-0.297	-0.342
b15	-0.562	-0.709	-0.380	b38	-0.348	-0.053	-0.300
b16	0.505	0.343	-0.380	b39	-1.202	-0.739	-0.380
b17	1.358	2.587	2.151	b40	-0.455	1.517	-0.304
b18	-1.202	0.103	-0.380	b41	-0.668	-0.489	-0.361
b19	-1.735	-0.709	-0.380	b42	1.038	2.766	-0.380
b20	-1.095	-0.581	-0.380	b43	-0.028	0.038	0.509
b21	-0.668	2.699	-0.380	b44	1.678	0.389	-0.281
b22	-1.095	-0.709	-0.380	b45	0.292	0.238	-0.323
b23	0.611	0.108	-0.342				

Appendix E

Constant Item Anchor Selection Results.

This appendix contains tables for the mean DIF effect for each item when a different anchor item was used as the reference item, following the constant item anchor selection method discussed in the present study.

Table E1.

Mean DIF Effect for EPvEB – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	1.064	m22	1.170
m02	0.946	m23	0.908
m03	0.894	m24	1.988
m04	0.928	m25	0.884
m05	1.033	m26	1.020
m06	0.979	m27	0.892
m07	0.884	m28	1.296
m08	0.890	m29	1.126
m09	2.714	m30	3.605
m10	0.963	m31	1.042
m11	0.998	m32	0.916
m12	1.791	m33	1.579
m13	0.892	m34	1.187
m14	3.612	m35	4.445
m15	1.007	m36	0.887
m16	1.585	m37	1.297
m17	0.909	m38	0.985
m18	1.114	m39	0.963
m19	1.299	m40	1.645
m20	1.362	m41	0.884
m21	0.965	m42	0.884

Table E2.

Mean DIF Effect for EPvSTEB – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	1.087	m22	1.249
m02	0.947	m23	0.911
m03	0.908	m24	1.975
m04	0.945	m25	0.909
m05	1.037	m26	1.045
m06	0.979	m27	0.910
m07	0.904	m28	1.252
m08	0.909	m29	1.107
m09	2.723	m30	3.689
m10	0.954	m31	1.071
m11	1.002	m32	0.934
m12	1.782	m33	1.577
m13	0.911	m34	1.195
m14	3.673	m35	4.413
m15	0.993	m36	0.903
m16	1.524	m37	1.284
m17	0.942	m38	0.990
m18	1.149	m39	0.967
m19	1.231	m40	1.650
m20	1.420	m41	0.903
m21	0.983	m42	0.904

Table E3.

Mean DIF Effect for EPvLTEB – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	1.023	m22	1.005
m02	0.942	m23	0.927
m03	0.886	m24	2.023
m04	0.897	m25	0.933
m05	1.030	m26	0.979
m06	0.990	m27	0.886
m07	0.911	m28	1.405
m08	0.890	m29	1.180
m09	2.693	m30	3.432
m10	1.001	m31	0.986
m11	0.997	m32	0.896
m12	1.811	m33	1.588
m13	0.888	m34	1.182
m14	3.484	m35	4.514
m15	1.054	m36	0.886
m16	1.736	m37	1.334
m17	0.888	m38	0.980
m18	1.045	m39	0.964
m19	1.465	m40	1.637
m20	1.242	m41	0.904
m21	0.941	m42	0.899

Table E4.

Mean DIF Effect for STEBvLTEB – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	0.151	m22	0.300
m02	0.140	m23	0.170
m03	0.159	m24	0.147
m04	0.205	m25	0.414
m05	0.142	m26	0.157
m06	0.148	m27	0.213
m07	0.280	m28	0.198
m08	0.227	m29	0.174
m09	0.140	m30	0.286
m10	0.177	m31	0.251
m11	0.140	m32	0.182
m12	0.150	m33	0.142
m13	0.224	m34	0.142
m14	0.223	m35	0.164
m15	0.167	m36	0.150
m16	0.309	m37	0.151
m17	0.310	m38	0.140
m18	0.257	m39	0.140
m19	0.375	m40	0.140
m20	0.224	m41	0.238
m21	0.156	m42	0.238

Table E5.

Mean DIF Effect for EPvSPA – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	1.137	m22	1.354
m02	1.085	m23	0.976
m03	0.958	m24	2.163
m04	1.024	m25	0.956
m05	1.131	m26	1.118
m06	1.040	m27	0.981
m07	0.956	m28	1.366
m08	0.955	m29	1.175
m09	3.101	m30	3.738
m10	1.055	m31	1.113
m11	1.111	m32	0.966
m12	1.754	m33	1.746
m13	0.972	m34	1.267
m14	3.870	m35	4.832
m15	1.164	m36	0.965
m16	1.767	m37	1.337
m17	0.955	m38	1.031
m18	1.200	m39	1.098
m19	1.323	m40	1.746
m20	1.488	m41	0.957
m21	1.021	m42	0.955

Table E6.

Mean DIF Effect for EPvOTH – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	0.978	m22	0.921
m02	0.798	m23	0.827
m03	0.810	m24	1.735
m04	0.812	m25	0.822
m05	0.912	m26	0.901
m06	0.934	m27	0.800
m07	0.804	m28	1.204
m08	0.807	m29	1.060
m09	2.166	m30	3.416
m10	0.858	m31	0.943
m11	0.866	m32	0.857
m12	1.846	m33	1.333
m13	0.801	m34	1.088
m14	3.242	m35	3.991
m15	0.840	m36	0.799
m16	1.311	m37	1.237
m17	0.867	m38	0.969
m18	1.010	m39	0.823
m19	1.263	m40	1.490
m20	1.191	m41	0.798
m21	0.923	m42	0.823

Table E7.

Mean DIF Effect for OTHvSPA – Mathematics Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
m01	0.237	m22	0.474
m02	0.490	m23	0.227
m03	0.228	m24	0.427
m04	0.281	m25	0.354
m05	0.303	m26	0.267
m06	0.232	m27	0.256
m07	0.275	m28	0.240
m08	0.227	m29	0.227
m09	0.963	m30	0.369
m10	0.251	m31	0.240
m11	0.294	m32	0.256
m12	0.307	m33	0.447
m13	0.230	m34	0.245
m14	0.631	m35	0.825
m15	0.401	m36	0.227
m16	0.496	m37	0.232
m17	0.291	m38	0.249
m18	0.265	m39	0.372
m19	0.255	m40	0.300
m20	0.313	m41	0.236
m21	0.240	m42	0.344

Table E8.

Mean DIF Effect for EPvEB – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.747	b24	0.685
b02	0.672	b25	0.965
b03	0.738	b26	0.711
b04	0.673	b27	0.677
b05	0.736	b28	1.158
b06	0.740	b29	0.674
b07	0.883	b30	0.875
b08	0.733	b31	0.674
b09	0.710	b32	3.031
b10	0.694	b33	0.911
b11	0.848	b34	0.722
b12	3.869	b35	0.772
b13	0.751	b36	0.883
b14	0.680	b37	0.700
b15	0.672	b38	0.672
b16	0.674	b39	0.846
b17	0.969	b40	0.724
b18	0.674	b41	0.767
b19	0.802	b42	0.908
b20	0.810	b43	0.860
b21	0.964	b44	3.091
b22	0.856	b45	3.125
b23	2.996		

Table E9.

Mean DIF Effect for EPvSTEB – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.759	b24	0.692
b02	0.683	b25	1.058
b03	0.745	b26	0.750
b04	0.682	b27	0.688
b05	0.735	b28	1.261
b06	0.746	b29	0.684
b07	0.949	b30	0.919
b08	0.761	b31	0.689
b09	0.722	b32	2.952
b10	0.695	b33	0.852
b11	0.850	b34	0.766
b12	4.037	b35	0.792
b13	0.772	b36	0.925
b14	0.685	b37	0.693
b15	0.689	b38	0.685
b16	0.682	b39	0.881
b17	0.953	b40	0.738
b18	0.683	b41	0.768
b19	0.904	b42	0.938
b20	0.839	b43	0.935
b21	0.971	b44	3.215
b22	0.836	b45	2.978
b23	3.015		

Table E10.

Mean DIF Effect for EPvLTEB – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.724	b24	0.678
b02	0.661	b25	0.742
b03	0.743	b26	0.657
b04	0.694	b27	0.667
b05	0.779	b28	0.893
b06	0.748	b29	0.720
b07	0.747	b30	0.789
b08	0.680	b31	0.657
b09	0.697	b32	3.261
b10	0.716	b33	1.137
b11	0.857	b34	0.660
b12	3.458	b35	0.744
b13	0.704	b36	0.779
b14	0.691	b37	0.772
b15	0.661	b38	0.658
b16	0.675	b39	0.783
b17	1.028	b40	0.707
b18	0.666	b41	0.803
b19	0.658	b42	0.834
b20	0.754	b43	0.716
b21	0.962	b44	2.842
b22	0.976	b45	3.548
b23	2.971		

Table E11.

Mean DIF Effect for STEBvLTEB – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.186	b24	0.205
b02	0.193	b25	0.427
b03	0.183	b26	0.340
b04	0.253	b27	0.183
b05	0.211	b28	0.465
b06	0.185	b29	0.364
b07	0.309	b30	0.245
b08	0.200	b31	0.207
b09	0.194	b32	0.364
b10	0.219	b33	0.337
b11	0.199	b34	0.282
b12	0.578	b35	0.190
b13	0.183	b36	0.228
b14	0.209	b37	0.340
b15	0.202	b38	0.184
b16	0.233	b39	0.220
b17	0.186	b40	0.188
b18	0.183	b41	0.193
b19	0.528	b42	0.196
b20	0.214	b43	0.348
b21	0.186	b44	0.436
b22	0.231	b45	0.623
b23	0.184		

Table E12.

Mean DIF Effect for EPvSPA – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.795	b24	0.735
b02	0.723	b25	1.067
b03	0.822	b26	0.759
b04	0.746	b27	0.731
b05	0.815	b28	1.218
b06	0.799	b29	0.732
b07	0.968	b30	0.890
b08	0.810	b31	0.723
b09	0.733	b32	3.271
b10	0.751	b33	1.011
b11	0.941	b34	0.738
b12	3.988	b35	0.821
b13	0.845	b36	0.944
b14	0.725	b37	0.741
b15	0.723	b38	0.725
b16	0.724	b39	0.910
b17	1.135	b40	0.777
b18	0.728	b41	0.849
b19	0.869	b42	1.012
b20	0.906	b43	0.918
b21	1.029	b44	3.256
b22	0.910	b45	3.594
b23	3.264		

Table E13.

Mean DIF Effect for EPvOTH – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.683	b24	0.606
b02	0.601	b25	0.804
b03	0.613	b26	0.638
b04	0.609	b27	0.603
b05	0.617	b28	1.063
b06	0.652	b29	0.598
b07	0.755	b30	0.887
b08	0.627	b31	0.620
b09	0.695	b32	2.653
b10	0.607	b33	0.761
b11	0.701	b34	0.733
b12	3.688	b35	0.700
b13	0.627	b36	0.788
b14	0.630	b37	0.638
b15	0.599	b38	0.598
b16	0.601	b39	0.746
b17	0.728	b40	0.656
b18	0.598	b41	0.637
b19	0.696	b42	0.743
b20	0.659	b43	0.772
b21	0.867	b44	2.849
b22	0.777	b45	2.333
b23	2.593		

Table E14.

Mean DIF Effect for OTHvSPA – Biology Assessment

Item	Mean DIF Effect	Item	Mean DIF Effect
b01	0.193	b24	0.191
b02	0.192	b25	0.266
b03	0.276	b26	0.195
b04	0.304	b27	0.192
b05	0.252	b28	0.199
b06	0.192	b29	0.236
b07	0.245	b30	0.207
b08	0.222	b31	0.262
b09	0.298	b32	0.646
b10	0.190	b33	0.266
b11	0.252	b34	0.346
b12	0.373	b35	0.189
b13	0.257	b36	0.195
b14	0.256	b37	0.201
b15	0.189	b38	0.210
b16	0.188	b39	0.213
b17	0.419	b40	0.190
b18	0.189	b41	0.260
b19	0.216	b42	0.275
b20	0.297	b43	0.202
b21	0.213	b44	0.449
b22	0.196	b45	1.366
b23	0.705		

Appendix F

Item Difficulties by Comparison Group.

This appendix contains tables for the item difficulties and thresholds for each comparison group, with units in logits. These tables also include the item difficulties and thresholds for these reference and focal groups in the "base model" as described in methods section. The differences between the reference and focal groups' item difficulties and thresholds are listed. The reference item is marked with an asterisk (*). The item difficulties and threshold difficulties for anchor set items represent the average difference in item difficulty between groups, and the average difference in item thresholds between groups, for the polytomous items. The bottom of each table lists the correlations between factor scores and the item difficulty or threshold for full credit on the item.

Table F1.

EPvEB Item Difficulties and Thresholds – Mathematics Assessment

Item	EP	EB	Difference	Item	EP	EB	Difference
m01	-1.812	-0.726	1.086	m22	-0.633	0.314	0.947
m02	-0.459	1.436	1.894	m23	-1.615	-0.195	1.420
m03	-0.738	0.756	1.494	m24	0.113	0.183	0.070
m04	-0.385	1.450	1.834	m25	-1.609	0.000	1.608
m05	-0.316	1.841	2.157	m26	-0.749	0.400	1.148
m06	-1.701	-0.485	1.216	m27	-1.082	0.600	1.682
m07	0.712	2.296	1.585	m28	-0.092	0.708	0.800
m08	-0.313	1.206	1.519	m29	-0.488	1.886	2.374
m09.1	-4.513	0.068	4.580	m30.1	-5.692	-0.072	5.620
m09.2	0.050	3.602	3.551	m30.2	-1.129	3.462	4.591
m09.3	2.358	5.732	3.373	m30.3	1.179	5.592	4.413
m09.4	3.579	7.102	3.522	m30.4	2.400	6.962	4.562
m10	-0.517	0.735	1.252	m31	-1.648	0.532	2.180
m11	-0.408	0.775	1.183	m32	0.127	1.518	1.391
m12.1	-2.816	0.624	3.440	m33.1	-4.868	-1.713	3.155
m12.2	1.747	4.158	2.411	m33.2	-0.305	1.821	2.126
m13*	-0.665	1.016	1.681	m34	-0.347	0.580	0.927
m14.1	-5.614	0.014	5.628	m35.1	-5.348	1.154	6.502
m14.2	-1.051	3.548	4.599	m35.2	-0.785	4.688	5.473
m14.3	1.257	5.678	4.421	m35.3	1.523	6.818	5.295
m14.4	2.478	7.048	4.569	m35.4	2.744	8.188	5.444
m15	0.127	1.295	1.168	m36	-1.085	0.563	1.647
m16.1	-3.930	-0.765	3.165	m37.1	-2.089	0.610	2.699
m16.2	0.633	2.769	2.136	m37.2	2.474	4.144	1.670
m17	-0.372	1.044	1.417	m38	-1.100	0.106	1.206
m18	0.058	2.407	2.348	m39	-0.276	0.975	1.251
m19.1	-2.248	0.455	2.703	m40.1	-3.832	-0.583	3.248
m19.2	2.315	3.989	1.674	m40.2	0.732	2.950	2.219
m20	0.089	0.816	0.726	m41	-0.372	1.218	1.591
m21	-0.552	0.693	1.246	m42	-1.167	0.449	1.616
LEX	0.614	0.714	0.694				
NP	0.222	0.363	0.468				
RC	0.560	0.699	0.732				

Table F2.

EPvSTEB Item Difficulties and Thresholds – Mathematics Assessment

Item	EP	STEB	Difference	Item	EP	STEB	Difference
m01	-1.813	-0.788	1.025	m22	-0.633	0.183	0.816
m02	-0.459	1.402	1.861	m23	-1.616	-0.201	1.415
m03	-0.738	0.691	1.429	m24	0.113	0.158	0.045
m04	-0.385	1.470	1.855	m25	-1.609	0.094	1.703
m05	-0.316	1.814	2.130	m26	-0.749	0.335	1.084
m06	-1.701	-0.510	1.191	m27	-1.082	0.621	1.704
m07	0.712	2.187	1.474	m28	-0.092	0.721	0.813
m08	-0.313	1.108	1.421	m29	-0.488	1.817	2.305
m09.1	-4.514	0.033	4.546	m30.1	-5.694	-0.021	5.673
m09.2	0.051	3.586	3.536	m30.2	-1.129	3.533	4.662
m09.3	2.359	5.669	3.310	m30.3	1.179	5.616	4.436
m09.4	3.580	6.954	3.374	m30.4	2.401	6.901	4.501
m10	-0.517	0.736	1.253	m31	-1.649	0.571	2.219
m11	-0.408	0.740	1.148	m32	0.127	1.443	1.316
m12.1	-2.817	0.567	3.384	m33.1	-4.870	-1.761	3.109
m12.2	1.748	4.121	2.373	m33.2	-0.305	1.793	2.098
m13*	-0.666	1.044	1.709	m34	-0.347	0.536	0.884
m14.1	-5.616	0.039	5.655	m35.1	-5.349	1.084	6.433
m14.2	-1.051	3.593	4.644	m35.2	-0.785	4.638	5.422
m14.3	1.258	5.676	4.418	m35.3	1.524	6.720	5.196
m14.4	2.479	6.961	4.482	m35.4	2.745	8.006	5.260
m15	0.128	1.290	1.163	m36	-1.085	0.541	1.626
m16.1	-3.931	-0.902	3.029	m37.1	-2.090	0.551	2.641
m16.2	0.634	2.652	2.018	m37.2	2.474	4.105	1.630
m17	-0.372	0.915	1.287	m38	-1.100	0.068	1.168
m18	0.059	2.452	2.393	m39	-0.276	0.941	1.217
m19.1	-2.249	0.301	2.549	m40.1	-3.833	-0.622	3.210
m19.2	2.316	3.854	1.539	m40.2	0.732	2.932	2.200
m20	0.089	0.717	0.627	m41	-0.372	1.116	1.488
m21	-0.552	0.630	1.182	m42	-1.168	0.481	1.649
LEX	0.614	0.719	0.696				
NP	0.222	0.363	0.466				
RC	0.560	0.703	0.733				

Table F3.

EPvLTEB Item Difficulties and Thresholds – Mathematics Assessment

Item	EP	LTEB	Difference	Item	EP	LTEB	Difference
m01	-1.813	-0.592	1.222	m22	-0.633	0.613	1.246
m02	-0.459	1.512	1.971	m23	-1.616	-0.182	1.434
m03	-0.738	0.903	1.641	m24	0.113	0.237	0.123
m04	-0.385	1.410	1.795	m25	-1.609	-0.198	1.412
m05	-0.316	1.904	2.220	m26	-0.749	0.544	1.293
m06	-1.702	-0.431	1.270	m27	-1.083	0.556	1.639
m07	0.713	2.576	1.864	m28	-0.092	0.680	0.772
m08	-0.313	1.435	1.748	m29	-0.488	2.050	2.538
m09.1	-4.515	0.145	4.660	m30.1	-5.695	-0.173	5.521
m09.2	0.051	3.645	3.595	m30.2	-1.129	3.327	4.456
m09.3	2.360	5.928	3.568	m30.3	1.180	5.610	4.430
m09.4	3.581	7.667	4.086	m30.4	2.401	7.349	4.948
m10	-0.517	0.735	1.253	m31	-1.649	0.453	2.102
m11	-0.408	0.851	1.259	m32	0.127	1.693	1.566
m12.1	-2.818	0.752	3.569	m33.1	-4.871	-1.612	3.259
m12.2	1.748	4.252	2.504	m33.2	-0.305	1.888	2.194
m13*	-0.666	0.958	1.624	m34	-0.347	0.675	1.022
m14.1	-5.617	-0.038	5.578	m35.1	-5.350	1.310	6.660
m14.2	-1.051	3.462	4.513	m35.2	-0.784	4.810	5.595
m14.3	1.258	5.745	4.487	m35.3	1.525	7.093	5.568
m14.4	2.479	7.484	5.005	m35.4	2.746	8.832	6.086
m15	0.128	1.309	1.181	m36	-1.085	0.610	1.695
m16.1	-3.932	-0.461	3.471	m37.1	-2.091	0.741	2.832
m16.2	0.634	3.040	2.406	m37.2	2.475	4.241	1.767
m17	-0.372	1.351	1.724	m38	-1.100	0.189	1.289
m18	0.059	2.312	2.253	m39	-0.276	1.052	1.328
m19.1	-2.249	0.811	3.060	m40.1	-3.833	-0.501	3.333
m19.2	2.316	4.311	1.995	m40.2	0.732	3.000	2.268
m20	0.090	1.042	0.952	m41	-0.372	1.459	1.831
m21	-0.553	0.835	1.388	m42	-1.168	0.381	1.549
LEX	0.614	0.702	0.682				
NP	0.222	0.362	0.460				
RC	0.560	0.692	0.726				

Table F4.

STEBvLTEB Item Difficulties and Thresholds – Mathematics Assessment

Item	STEB	LTEB	Difference	Item	STEB	LTEB	Difference
m01	-0.774	-0.585	0.189	m22	0.182	0.603	0.421
m02	1.372	1.486	0.114	m23	-0.196	-0.180	0.016
m03	0.677	0.888	0.210	m24	0.157	0.233	0.076
m04	1.437	1.385	-0.051	m25	0.092	-0.197	-0.289
m05	1.773	1.872	0.098	m26	0.330	0.535	0.206
m06	-0.501	-0.427	0.074	m27	0.610	0.546	-0.063
m07	2.138	2.535	0.397	m28	0.708	0.670	-0.038
m08	1.085	1.411	0.326	m29	1.777	2.016	0.239
m09.1	0.041	0.147	0.106	m30.1	-0.009	-0.163	-0.154
m09.2	3.528	3.606	0.077	m30.2	3.479	3.296	-0.183
m09.3	5.595	5.879	0.284	m30.3	5.546	5.570	0.024
m09.4	6.879	7.617	0.739	m30.4	6.829	7.308	0.478
m10	0.722	0.723	0.001	m31	0.560	0.445	-0.115
m11	0.726	0.837	0.111	m32	1.413	1.666	0.253
m12.1	0.560	0.744	0.183	m33.1	-1.736	-1.595	0.141
m12.2	4.048	4.202	0.154	m33.2	1.752	1.864	0.112
m13*	1.023	0.942	-0.081	m34	0.527	0.665	0.138
m14.1	0.046	-0.033	-0.079	m35.1	1.075	1.295	0.220
m14.2	3.533	3.426	-0.108	m35.2	4.563	4.754	0.191
m14.3	5.600	5.699	0.099	m35.3	6.630	7.028	0.398
m14.4	6.884	7.437	0.554	m35.4	7.913	8.766	0.852
m15	1.264	1.287	0.023	m36	0.531	0.599	0.068
m16.1	-0.885	-0.453	0.432	m37.1	0.544	0.731	0.187
m16.2	2.603	3.006	0.403	m37.2	4.032	4.190	0.158
m17	0.896	1.328	0.432	m38	0.067	0.185	0.118
m18	2.395	2.273	-0.122	m39	0.923	1.034	0.111
m19.1	0.300	0.801	0.501	m40.1	-0.610	-0.492	0.117
m19.2	3.788	4.260	0.472	m40.2	2.878	2.966	0.088
m20	0.704	1.024	0.321	m41	1.093	1.435	0.342
m21	0.618	0.821	0.204	m42	0.473	0.374	-0.099
LEX	0.720	0.702	0.339				
NP	0.364	0.362	0.239				
RC	0.704	0.693	0.389				

Table F5.

EPvSPA Item Difficulties and Thresholds – Mathematics Assessment

Item	EP	SPA	Difference	Item	EP	SPA	Difference
m01	-1.813	-0.579	1.234	m22	-0.633	0.322	0.955
m02	-0.459	1.850	2.309	m23	-1.615	-0.023	1.593
m03	-0.738	0.924	1.662	m24	0.113	0.210	0.097
m04	-0.385	1.712	2.097	m25	-1.609	0.061	1.670
m05	-0.316	2.126	2.442	m26	-0.749	0.515	1.264
m06	-1.701	-0.291	1.410	m27	-1.082	0.835	1.917
m07	0.712	2.387	1.675	m28	-0.092	0.848	0.941
m08	-0.313	1.378	1.691	m29	-0.488	2.055	2.543
m09.1	-4.513	0.679	5.192	m30.1	-5.693	0.243	5.936
m09.2	0.051	4.184	4.134	m30.2	-1.129	3.748	4.877
m09.3	2.359	6.599	4.240	m30.3	1.179	6.163	4.984
m09.4	3.580	8.430	4.850	m30.4	2.400	7.994	5.594
m10	-0.517	0.863	1.380	m31	-1.648	0.744	2.393
m11	-0.408	0.867	1.275	m32	0.127	1.755	1.628
m12.1	-2.817	0.710	3.527	m33.1	-4.869	-1.353	3.516
m12.2	1.747	4.215	2.468	m33.2	-0.305	2.153	2.458
m13*	-0.666	1.203	1.868	m34	-0.347	0.714	1.062
m14.1	-5.615	0.467	6.082	m35.1	-5.349	1.743	7.091
m14.2	-1.051	3.972	5.023	m35.2	-0.785	5.248	6.032
m14.3	1.257	6.387	5.129	m35.3	1.524	7.662	6.139
m14.4	2.479	8.218	5.739	m35.4	2.745	9.494	6.749
m15	0.127	1.324	1.196	m36	-1.085	0.736	1.821
m16.1	-3.930	-0.386	3.544	m37.1	-2.090	0.766	2.856
m16.2	0.633	3.119	2.486	m37.2	2.474	4.272	1.798
m17	-0.372	1.314	1.686	m38	-1.100	0.331	1.431
m18	0.059	2.655	2.596	m39	-0.276	1.021	1.297
m19.1	-2.249	0.582	2.831	m40.1	-3.832	-0.315	3.517
m19.2	2.315	4.088	1.772	m40.2	0.732	3.190	2.458
m20	0.089	0.895	0.806	m41	-0.372	1.362	1.734
m21	-0.552	0.905	1.457	m42	-1.168	0.513	1.681
LEX	0.614	0.715	0.689				
NP	0.222	0.371	0.454				
RC	0.560	0.697	0.711				

Table F6.

EPvOTH Item Difficulties and Thresholds – Mathematics Assessment

Item	EP	OTH	Difference	Item	EP	OTH	Difference
m01	-1.813	-0.958	0.855	m22	-0.633	0.304	0.937
m02	-0.459	0.913	1.372	m23	-1.616	-0.459	1.157
m03	-0.738	0.503	1.241	m24	0.113	0.141	0.028
m04	-0.385	1.088	1.472	m25	-1.610	-0.096	1.514
m05	-0.316	1.467	1.783	m26	-0.749	0.222	0.971
m06	-1.702	-0.786	0.916	m27	-1.083	0.255	1.337
m07	0.713	2.139	1.426	m28	-0.092	0.493	0.585
m08	-0.313	0.950	1.263	m29	-0.488	1.638	2.126
m09.1	-4.515	-0.813	3.702	m30.1	-5.695	-0.545	5.150
m09.2	0.051	2.799	2.748	m30.2	-1.129	3.067	4.196
m09.3	2.360	4.797	2.437	m30.3	1.180	5.065	3.885
m09.4	3.581	6.023	2.442	m30.4	2.401	6.291	3.890
m10	-0.517	0.539	1.057	m31	-1.649	0.216	1.866
m11	-0.408	0.630	1.038	m32	0.127	1.186	1.058
m12.1	-2.818	0.489	3.307	m33.1	-4.871	-2.252	2.619
m12.2	1.748	4.100	2.352	m33.2	-0.305	1.360	1.665
m13*	-0.666	0.738	1.404	m34	-0.347	0.373	0.720
m14.1	-5.617	-0.659	4.958	m35.1	-5.350	0.403	5.753
m14.2	-1.051	2.953	4.004	m35.2	-0.784	4.014	4.799
m14.3	1.258	4.951	3.693	m35.3	1.525	6.013	4.488
m14.4	2.480	6.177	3.698	m35.4	2.746	7.239	4.493
m15	0.128	1.240	1.112	m36	-1.085	0.301	1.386
m16.1	-3.932	-1.344	2.587	m37.1	-2.091	0.373	2.464
m16.2	0.634	2.267	1.633	m37.2	2.475	3.984	1.509
m17	-0.373	0.663	1.036	m38	-1.100	-0.233	0.867
m18	0.059	2.080	2.022	m39	-0.276	0.898	1.174
m19.1	-2.250	0.261	2.510	m40.1	-3.833	-0.994	2.840
m19.2	2.316	3.872	1.556	m40.2	0.732	2.618	1.885
m20	0.090	0.689	0.600	m41	-0.372	0.999	1.372
m21	-0.553	0.380	0.932	m42	-1.168	0.347	1.515
LEX	0.614	0.710	0.693				
NP	0.222	0.349	0.464				
RC	0.560	0.707	0.771				

Table F7.

OTHvSPA Item Difficulties and Thresholds – Mathematics Assessment

Item	OTH	SPA	Difference	Item	OTH	SPA	Difference
m01	-0.936	-0.572	0.364	m22	0.298	0.317	0.019
m02	0.888	1.819	0.931	m23	-0.448	-0.024	0.424
m03	0.490	0.909	0.419	m24	0.140	0.207	0.068
m04	1.056	1.682	0.627	m25	-0.094	0.058	0.152
m05	1.424	2.090	0.666	m26	0.218	0.507	0.289
m06	-0.768	-0.288	0.479	m27	0.249	0.821	0.572
m07	2.077	2.349	0.272	m28	0.482	0.835	0.353
m08	0.924	1.355	0.431	m29	1.591	2.021	0.430
m09.1	-0.784	0.671	1.455	m30.1	-0.518	0.247	0.765
m09.2	2.743	4.136	1.394	m30.2	3.009	3.712	0.703
m09.3	4.720	6.542	1.822	m30.3	4.986	6.117	1.131
m09.4	5.943	8.372	2.429	m30.4	6.209	7.947	1.738
m10	0.526	0.849	0.323	m31	0.212	0.732	0.520
m11	0.614	0.852	0.238	m32	1.153	1.726	0.573
m12.1	0.483	0.702	0.218	m33.1	-2.213	-1.337	0.875
m12.2	4.010	4.167	0.156	m33.2	1.314	2.128	0.813
m13*	0.719	1.183	0.463	m34	0.365	0.703	0.339
m14.1	-0.634	0.462	1.096	m35.1	0.410	1.721	1.311
m14.2	2.893	3.927	1.034	m35.2	3.937	5.186	1.249
m14.3	4.870	6.332	1.462	m35.3	5.914	7.591	1.677
m14.4	6.093	8.163	2.069	m35.4	7.137	9.421	2.284
m15	1.206	1.302	0.096	m36	0.294	0.723	0.430
m16.1	-1.317	-0.378	0.939	m37.1	0.370	0.755	0.386
m16.2	2.210	3.087	0.877	m37.2	3.896	4.220	0.324
m17	0.646	1.292	0.646	m38	-0.226	0.325	0.551
m18	2.018	2.611	0.593	m39	0.874	1.005	0.130
m19.1	0.262	0.575	0.313	m40.1	-0.971	-0.309	0.662
m19.2	3.789	4.040	0.251	m40.2	2.556	3.155	0.600
m20	0.672	0.881	0.209	m41	0.972	1.339	0.367
m21	0.370	0.890	0.520	m42	0.339	0.504	0.165
LEX	0.710	0.715	0.597				
NP	0.349	0.372	0.380				
RC	0.709	0.698	0.528				

Table F8.

EPvEB Item Difficulties and Thresholds – Biology Assessment

Item	EP	EB	Difference	Item	EP	EB	Difference
b01	-1.366	0.204	1.570	b25	-1.472	0.461	1.933
b02*	-0.997	0.310	1.307	b26	-0.534	0.536	1.070
b03	-0.653	0.334	0.988	b27	-1.374	-0.175	1.199
b04	-0.918	0.311	1.229	b28	-1.669	0.527	2.196
b05	-0.519	0.474	0.993	b29	-1.488	-0.165	1.323
b06	-1.073	-0.089	0.984	b30	-0.799	-0.046	0.753
b07	-0.937	-0.194	0.744	b31	-1.102	0.116	1.219
b08	-1.949	-0.409	1.539	b32.0	-4.300	0.301	4.601
b09	-1.002	0.474	1.477	b32.1	1.058	4.520	3.463
b10	-0.088	1.041	1.129	b32.2	2.677	6.335	3.659
b11	-1.293	0.459	1.752	b32.3	4.020	7.518	3.498
b12.0	-4.934	0.552	5.486	b33	0.251	0.963	0.712
b12.1	0.424	4.772	4.348	b34	-1.402	0.111	1.514
b12.2	2.043	6.587	4.544	b35	-0.406	0.510	0.916
b12.3	3.386	7.770	4.383	b36	-1.231	0.579	1.811
b13	-1.300	0.277	1.577	b37	-1.332	0.110	1.442
b14	-1.264	-0.078	1.186	b38	-0.947	0.350	1.298
b15	-0.985	0.326	1.310	b39	-1.060	-0.268	0.793
b16	-0.834	0.495	1.329	b40	-1.305	0.214	1.519
b17	0.339	0.987	0.648	b41	-0.718	0.209	0.927
b18	-0.883	0.339	1.223	b42	-1.533	0.316	1.849
b19	-0.646	0.216	0.862	b43	-0.642	0.132	0.774
b20	-0.480	0.370	0.850	b44.0	-4.431	0.239	4.670
b21	-0.317	0.336	0.653	b44.1	0.926	4.458	3.532
b22	-0.110	0.668	0.778	b44.2	2.546	6.273	3.728
b23.0	-4.313	0.245	4.558	b44.3	3.889	7.456	3.568
b23.1	1.045	4.465	3.420	b45.0	-6.109	-1.401	4.708
b23.2	2.664	6.280	3.616	b45.1	-0.752	2.818	3.570
b23.3	4.007	7.463	3.456	b45.2	0.868	4.633	3.766
b24	-1.225	0.158	1.383	b45.3	2.211	5.816	3.605
LEX	0.331	0.379	0.397				
NP	0.094	0.130	0.170				
RC	-0.135	-0.087	0.011				

Table F9.

EPvSTEB Item Difficulties and Thresholds – Biology Assessment

Item	EP	STEB	Difference	Item	EP	STEB	Difference
b01	-1.367	0.186	1.553	b25	-1.473	0.567	2.039
b02*	-0.998	0.282	1.280	b26	-0.534	0.443	0.977
b03	-0.654	0.335	0.989	b27	-1.375	-0.185	1.190
b04	-0.918	0.354	1.272	b28	-1.670	0.645	2.315
b05	-0.520	0.498	1.018	b29	-1.489	-0.270	1.219
b06	-1.074	-0.086	0.988	b30	-0.800	-0.104	0.695
b07	-0.938	-0.279	0.659	b31	-1.103	0.077	1.180
b08	-1.950	-0.394	1.556	b32.0	-4.302	0.187	4.489
b09	-1.003	0.446	1.449	b32.1	1.059	4.427	3.368
b10	-0.088	1.071	1.159	b32.2	2.679	6.150	3.471
b11	-1.294	0.426	1.720	b32.3	4.022	7.335	3.312
b12.0	-4.936	0.708	5.644	b33	0.251	1.039	0.788
b12.1	0.425	4.948	4.523	b34	-1.403	0.163	1.567
b12.2	2.045	6.671	4.626	b35	-0.407	0.485	0.892
b12.3	3.388	7.856	4.467	b36	-1.232	0.612	1.845
b13	-1.301	0.278	1.579	b37	-1.333	0.014	1.347
b14	-1.265	-0.057	1.209	b38	-0.948	0.353	1.301
b15	-0.985	0.345	1.330	b39	-1.061	-0.315	0.746
b16	-0.835	0.444	1.278	b40	-1.306	0.191	1.497
b17	0.340	0.995	0.655	b41	-0.719	0.220	0.939
b18	-0.884	0.337	1.221	b42	-1.534	0.331	1.865
b19	-0.647	0.067	0.714	b43	-0.643	0.033	0.676
b20	-0.480	0.327	0.808	b44.0	-4.433	0.351	4.784
b21	-0.318	0.318	0.636	b44.1	0.927	4.591	3.663
b22	-0.110	0.703	0.813	b44.2	2.547	6.313	3.766
b23.0	-4.315	0.249	4.564	b44.3	3.891	7.498	3.607
b23.1	1.046	4.489	3.443	b45.0	-6.112	-1.590	4.522
b23.2	2.666	6.212	3.546	b45.1	-0.752	2.649	3.401
b23.3	4.009	7.397	3.387	b45.2	0.868	4.372	3.504
b24	-1.226	0.120	1.346	b45.3	2.212	5.557	3.345
LEX	0.331	0.386	0.410				
NP	0.094	0.135	0.181				
RC	-0.135	-0.079	0.033				

Table F10.

EPvLTEB Item Difficulties and Thresholds – Biology Assessment

Item	EP	LTEB	Difference	Item	EP	LTEB	Difference
b01	-1.368	0.256	1.624	b25	-1.474	0.187	1.660
b02*	-0.998	0.387	1.386	b26	-0.535	0.804	1.339
b03	-0.654	0.333	0.987	b27	-1.376	-0.149	1.227
b04	-0.919	0.197	1.116	b28	-1.671	0.223	1.894
b05	-0.520	0.408	0.928	b29	-1.490	0.125	1.615
b06	-1.075	-0.097	0.977	b30	-0.800	0.115	0.915
b07	-0.939	0.041	0.979	b31	-1.104	0.224	1.328
b08	-1.951	-0.455	1.496	b32.0	-4.304	0.587	4.890
b09	-1.004	0.553	1.557	b32.1	1.059	4.782	3.722
b10	-0.088	0.963	1.051	b32.2	2.680	7.008	4.328
b11	-1.294	0.550	1.845	b32.3	4.024	8.201	4.176
b12.0	-4.938	0.169	5.107	b33	0.252	0.768	0.516
b12.1	0.425	4.364	3.938	b34	-1.404	-0.030	1.374
b12.2	2.046	6.590	4.544	b35	-0.407	0.577	0.984
b12.3	3.390	7.783	4.393	b36	-1.233	0.492	1.726
b13	-1.302	0.275	1.577	b37	-1.334	0.379	1.713
b14	-1.266	-0.139	1.127	b38	-0.949	0.344	1.292
b15	-0.986	0.275	1.261	b39	-1.062	-0.138	0.923
b16	-0.835	0.639	1.475	b40	-1.307	0.278	1.585
b17	0.340	0.970	0.630	b41	-0.719	0.177	0.896
b18	-0.885	0.346	1.230	b42	-1.535	0.276	1.812
b19	-0.647	0.644	1.291	b43	-0.643	0.408	1.051
b20	-0.480	0.488	0.968	b44.0	-4.435	-0.036	4.400
b21	-0.318	0.385	0.703	b44.1	0.928	4.159	3.231
b22	-0.110	0.577	0.688	b44.2	2.549	6.386	3.837
b23.0	-4.317	0.239	4.556	b44.3	3.893	7.578	3.685
b23.1	1.046	4.434	3.387	b45.0	-6.115	-0.913	5.201
b23.2	2.667	6.660	3.993	b45.1	-0.751	3.282	4.033
b23.3	4.011	7.853	3.842	b45.2	0.869	5.508	4.639
b24	-1.227	0.263	1.489	b45.3	2.213	6.700	4.487
LEX	0.331	0.359	0.351				
NP	0.094	0.117	0.136				
RC	-0.135	-0.107	-0.049				

Table F11.

STEBvLTEB Item Difficulties and Thresholds – Biology Assessment

Item	STEB	LTEB	Difference	Item	STEB	LTEB	Difference
b01	0.185	0.253	0.068	b25	0.559	0.185	-0.374
b02*	0.280	0.383	0.103	b26	0.437	0.794	0.357
b03	0.332	0.329	-0.003	b27	-0.180	-0.147	0.033
b04	0.350	0.196	-0.154	b28	0.636	0.220	-0.415
b05	0.492	0.404	-0.088	b29	-0.264	0.124	0.387
b06	-0.082	-0.096	-0.014	b30	-0.100	0.114	0.214
b07	-0.272	0.041	0.313	b31	0.078	0.221	0.143
b08	-0.386	-0.450	-0.064	b32.0	0.195	0.583	0.388
b09	0.440	0.547	0.107	b32.1	4.378	4.749	0.371
b10	1.053	0.951	-0.102	b32.2	6.095	6.973	0.878
b11	0.421	0.544	0.123	b32.3	7.277	8.165	0.888
b12.0	0.706	0.172	-0.534	b33	1.022	0.758	-0.263
b12.1	4.889	4.338	-0.551	b34	0.163	-0.030	-0.193
b12.2	6.606	6.562	-0.044	b35	0.479	0.570	0.091
b12.3	7.787	7.754	-0.034	b36	0.603	0.486	-0.117
b13	0.275	0.271	-0.004	b37	0.016	0.375	0.358
b14	-0.054	-0.137	-0.084	b38	0.349	0.339	-0.010
b15	0.341	0.271	-0.070	b39	-0.308	-0.137	0.171
b16	0.438	0.631	0.193	b40	0.190	0.274	0.084
b17	0.979	0.958	-0.020	b41	0.219	0.175	-0.044
b18	0.334	0.342	0.008	b42	0.327	0.273	-0.054
b19	0.068	0.636	0.568	b43	0.035	0.403	0.368
b20	0.324	0.483	0.158	b44.0	0.351	-0.033	-0.384
b21	0.315	0.381	0.066	b44.1	4.534	4.133	-0.402
b22	0.692	0.571	-0.121	b44.2	6.251	6.357	0.106
b23.0	0.251	0.238	-0.013	b44.3	7.433	7.548	0.116
b23.1	4.434	4.404	-0.030	b45.0	-1.561	-0.894	0.667
b23.2	6.151	6.628	0.477	b45.1	2.622	3.271	0.649
b23.3	7.333	7.820	0.487	b45.2	4.339	5.496	1.157
b24	0.121	0.260	0.139	b45.3	5.520	6.687	1.167
LEX	0.386	0.359	-0.037				
NP	0.134	0.117	-0.080				
RC	-0.079	-0.107	-0.263				

Table F12.

EPvSPA Item Difficulties and Thresholds – Biology Assessment

Item	EP	SPA	Difference	Item	EP	SPA	Difference
b01	-1.367	0.320	1.686	b25	-1.472	0.677	2.149
b02*	-0.997	0.426	1.424	b26	-0.534	0.689	1.223
b03	-0.653	0.389	1.042	b27	-1.375	-0.028	1.346
b04	-0.918	0.352	1.270	b28	-1.670	0.685	2.354
b05	-0.520	0.540	1.060	b29	-1.489	0.026	1.515
b06	-1.074	0.027	1.101	b30	-0.799	0.120	0.920
b07	-0.938	-0.120	0.818	b31	-1.103	0.327	1.430
b08	-1.949	-0.231	1.718	b32.0	-4.301	0.692	4.994
b09	-1.003	0.517	1.519	b32.1	1.058	4.950	3.892
b10	-0.088	1.161	1.249	b32.2	2.678	6.865	4.187
b11	-1.293	0.665	1.958	b32.3	4.021	8.078	4.057
b12.0	-4.935	0.824	5.760	b33	0.251	1.019	0.767
b12.1	0.424	5.082	4.658	b34	-1.403	0.135	1.538
b12.2	2.044	6.997	4.953	b35	-0.407	0.638	1.045
b12.3	3.388	8.210	4.823	b36	-1.232	0.732	1.964
b13	-1.301	0.486	1.787	b37	-1.333	0.215	1.548
b14	-1.265	0.128	1.393	b38	-0.948	0.521	1.469
b15	-0.985	0.459	1.444	b39	-1.061	-0.173	0.888
b16	-0.835	0.626	1.461	b40	-1.306	0.337	1.643
b17	0.339	0.977	0.637	b41	-0.718	0.272	0.990
b18	-0.884	0.477	1.361	b42	-1.534	0.536	2.069
b19	-0.647	0.308	0.955	b43	-0.642	0.236	0.878
b20	-0.480	0.414	0.894	b44.0	-4.433	0.543	4.976
b21	-0.318	0.430	0.747	b44.1	0.927	4.801	3.874
b22	-0.110	0.778	0.888	b44.2	2.547	6.716	4.169
b23.0	-4.314	0.670	4.985	b44.3	3.890	7.929	4.038
b23.1	1.045	4.928	3.883	b45.0	-6.111	-0.764	5.347
b23.2	2.665	6.843	4.178	b45.1	-0.752	3.494	4.246
b23.3	4.008	8.056	4.048	b45.2	0.868	5.409	4.541
b24	-1.225	0.302	1.527	b45.3	2.212	6.622	4.410
LEX	0.331	0.375	0.382				
NP	0.094	0.128	0.159				
RC	-0.135	-0.092	-0.013				

Table F13.

EPvOTH Item Difficulties and Thresholds – Biology Assessment

Item	EP	OTH	Difference	Item	EP	OTH	Difference
b01	-1.368	0.010	1.378	b25	-1.474	0.109	1.583
b02*	-0.999	0.114	1.112	b26	-0.535	0.283	0.818
b03	-0.654	0.242	0.896	b27	-1.376	-0.425	0.951
b04	-0.919	0.241	1.160	b28	-1.671	0.265	1.937
b05	-0.520	0.359	0.880	b29	-1.491	-0.490	1.000
b06	-1.075	-0.287	0.788	b30	-0.800	-0.327	0.473
b07	-0.939	-0.319	0.619	b31	-1.104	-0.236	0.868
b08	-1.951	-0.717	1.234	b32.0	-4.304	-0.350	3.954
b09	-1.004	0.400	1.404	b32.1	1.060	3.903	2.844
b10	-0.088	0.840	0.928	b32.2	2.680	5.692	3.012
b11	-1.295	0.122	1.416	b32.3	4.025	6.876	2.851
b12.0	-4.938	0.110	5.048	b33	0.252	0.864	0.612
b12.1	0.425	4.363	3.937	b34	-1.405	0.069	1.474
b12.2	2.046	6.152	4.106	b35	-0.407	0.294	0.701
b12.3	3.391	7.335	3.945	b36	-1.233	0.327	1.560
b13	-1.302	-0.068	1.234	b37	-1.334	-0.069	1.266
b14	-1.266	-0.427	0.839	b38	-0.949	0.067	1.015
b15	-0.986	0.101	1.087	b39	-1.062	-0.430	0.632
b16	-0.835	0.275	1.111	b40	-1.307	0.006	1.313
b17	0.340	0.998	0.658	b41	-0.719	0.101	0.821
b18	-0.885	0.108	0.992	b42	-1.535	-0.045	1.490
b19	-0.647	0.060	0.707	b43	-0.643	-0.045	0.598
b20	-0.480	0.294	0.775	b44.0	-4.436	-0.266	4.170
b21	-0.318	0.176	0.494	b44.1	0.928	3.987	3.059
b22	-0.110	0.482	0.592	b44.2	2.549	5.776	3.227
b23.0	-4.317	-0.432	3.885	b44.3	3.893	6.960	3.066
b23.1	1.047	3.821	2.774	b45.0	-6.115	-2.546	3.570
b23.2	2.667	5.610	2.943	b45.1	-0.751	1.707	2.459
b23.3	4.012	6.793	2.782	b45.2	0.869	3.496	2.627
b24	-1.227	-0.084	1.143	b45.3	2.214	4.680	2.466
LEX	0.331	0.385	0.422				
NP	0.094	0.133	0.188				
RC	-0.135	-0.079	0.051				

Table F14.

OTHvSPA Item Difficulties and Thresholds – Biology Assessment

Item	OTH	SPA	Difference	Item	OTH	SPA	Difference
b01	0.012	0.316	0.304	b25	0.109	0.667	0.558
b02*	0.114	0.421	0.307	b26	0.279	0.679	0.400
b03	0.240	0.384	0.144	b27	-0.414	-0.027	0.387
b04	0.238	0.348	0.110	b28	0.262	0.675	0.413
b05	0.354	0.533	0.179	b29	-0.478	0.026	0.504
b06	-0.278	0.028	0.306	b30	-0.317	0.120	0.437
b07	-0.310	-0.117	0.192	b31	-0.229	0.324	0.552
b08	-0.701	-0.227	0.474	b32.0	-0.328	0.688	1.015
b09	0.393	0.510	0.117	b32.1	3.854	4.908	1.054
b10	0.823	1.145	0.322	b32.2	5.633	6.820	1.187
b11	0.121	0.656	0.535	b32.3	6.811	8.031	1.221
b12.0	0.122	0.817	0.695	b33	0.846	1.004	0.158
b12.1	4.304	5.037	0.734	b34	0.071	0.134	0.063
b12.2	6.083	6.950	0.867	b35	0.290	0.629	0.340
b12.3	7.260	8.161	0.901	b36	0.322	0.721	0.399
b13	-0.064	0.479	0.544	b37	-0.064	0.213	0.277
b14	-0.416	0.127	0.543	b38	0.068	0.514	0.446
b15	0.101	0.453	0.352	b39	-0.419	-0.170	0.249
b16	0.272	0.618	0.346	b40	0.009	0.333	0.324
b17	0.976	0.963	-0.013	b41	0.102	0.269	0.167
b18	0.108	0.471	0.363	b42	-0.042	0.528	0.570
b19	0.061	0.304	0.244	b43	-0.042	0.233	0.275
b20	0.290	0.409	0.118	b44.0	-0.251	0.537	0.788
b21	0.175	0.425	0.250	b44.1	3.931	4.757	0.827
b22	0.474	0.767	0.294	b44.2	5.710	6.669	0.960
b23.0	-0.416	0.664	1.081	b44.3	6.887	7.881	0.994
b23.1	3.766	4.884	1.119	b45.0	-2.514	-0.741	1.773
b23.2	5.545	6.796	1.252	b45.1	1.668	3.479	1.811
b23.3	6.722	8.008	1.286	b45.2	3.447	5.391	1.944
b24	-0.079	0.298	0.378	b45.3	4.625	6.603	1.978
LEX	0.385	0.374	0.229				
NP	0.133	0.127	0.067				
RC	-0.080	-0.092	-0.142				

Appendix G

Rasch HGLM Results.

For each assessment and comparison group, there are two tables. The first table in the set contains HGLM model results and effect sizes for each significant DIF estimate based on the results from the 95% confidence intervals (CI) for adjusted DIF estimates in the second table in the set To aid readers in the detection and interpretation of changes in DIF between models, color was added to cells of the effect sizes of the item by EB status interaction parameter estimates in the first table of the set. Blue was used when the adjusted DIF estimate's 95% CI was above γ_{01}; this indicates the item was easier for the reference group, after controlling for ability and linguistic features (if applicable). Brown was used when the adjusted DIF estimate's 95% CI was below γ_{01}; this indicates the item was easier for the focal group, after controlling for ability and linguistic features (if applicable). Dark blue indicates substantial DIF favoring the reference group (ΔOR < 1.50), light blue for moderate DIF favoring the reference group (ΔOR > 1.50 and < 1.00), dark brown for substantial DIF favoring the focal group (ΔOR > 1.50), and light brown for moderate DIF favoring the focal group (ΔOR < 1.50 and > 1.00). For an item in a table, if the color changes from blue in the base model to brown in a model using a linguistic feature as a predictor (changes from favoring the reference group to favoring the focal group; the inverse is true if the color changes from brown to blue), this is evidence there are group differences in how test-takers respond to the item based on that linguistic feature in an item. Items that change from favoring the reference group to the focal group (or vice versa) are analyzed in the Results section.

Note: "+" indicates p < .10, "*" indicates p < .05, "**" indicates p < .01, and "***" indicates p < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial

DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G1.

EPvEB Model Results – Mathematics Assessment

Effect	Base model Estimate (SE)	Base model Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size
Intercept	-0.665*** (0.010)	-	3.935*** (0.176)	-
Intercept*EB Status	1.681*** (0.045)	-	-2.751*** (0.340)	-
NP	-	-	7.249*** (0.276)	-
NP*EB Status	-	-	-6.927*** (0.534)	-
m01	-1.147*** (0.013)	-	-1.26*** (0.013)	-
m01*EB Status	-0.595*** (0.052)	1.108	-0.466*** (0.052)	1.114
m02	0.207*** (0.013)	-	0.099*** (0.013)	-
m02*EB Status	0.213*** (0.058)	1.933	0.304*** (0.057)	1.900
m03	-0.072*** (0.013)	-	-2.011*** (0.075)	-
m03*EB Status	-0.187*** (0.054)	1.525	1.682*** (0.152)	1.518
m04	0.281*** (0.013)	-	0.273*** (0.012)	-
m04*EB Status	0.153** (0.059)	1.872	0.145* (0.058)	1.837
m05	0.349*** (0.013)	-	-5.445*** (0.222)	-
m05*EB Status	0.476*** (0.062)	2.201	6.068*** (0.431)	2.226
m06	-1.035*** (0.013)	-	-1.145*** (0.013)	-
m06*EB Status	-0.465*** (0.052)	1.241	-0.340*** (0.052)	1.243
m07	1.377*** (0.013)	-	1.53*** (0.015)	-
m07*EB Status	-0.097 (0.067)	1.617	-0.284*** (0.067)	1.597
m08	0.352*** (0.013)	-	0.545*** (0.015)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m08*EB Status	-0.162** (0.056)	1.550	-0.353*** (0.057)	1.526
m09	-3.847*** (0.012)	-	-3.819*** (0.015)	-
m09*EB Status	2.899*** (0.054)	4.674	2.913*** (0.055)	4.860
m10	0.148*** (0.013)	-	-8.185*** (0.318)	-
m10*EB Status	-0.429*** (0.054)	1.278	7.589*** (0.616)	1.304
m11	0.258*** (0.013)	-	0.454*** (0.015)	-
m11*EB Status	-0.499*** (0.054)	1.206	-0.680*** (0.055)	1.193
m12	-2.151*** (0.013)	-	-4.438*** (0.085)	-
m12*EB Status	1.759*** (0.054)	3.511	3.971*** (0.171)	3.585
m13	-	-	-	-
m13*EB Status	-	-	-	-
m14	-4.949*** (0.012)	-	-18.499*** (0.512)	-
m14*EB Status	3.947*** (0.053)	5.744	16.931*** (0.989)	5.875
m15	0.793*** (0.013)	-	-5.578*** (0.244)	-
m15*EB Status	-0.513*** (0.057)	1.192	5.627*** (0.474)	1.196
m16	-3.264*** (0.012)	-	-7.751*** (0.168)	-
m16*EB Status	1.484*** (0.052)	3.230	5.791*** (0.327)	3.322
m17	0.293*** (0.013)	-	0.184*** (0.013)	-
m17*EB Status	-0.264*** (0.056)	1.446	-0.161** (0.055)	1.425
m18	0.724*** (0.013)	-	0.701*** (0.012)	-
m18*EB Status	0.667*** (0.069)	2.396	0.643*** (0.067)	2.345
m19	-1.583*** (0.013)	-	-6.018*** (0.168)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m19*EB Status	1.022*** (0.053)	2.759	5.292*** (0.327)	2.812
m20	0.755*** (0.013)	-	0.430*** (0.017)	-
m20*EB Status	-0.955*** (0.054)	0.741	-0.635*** (0.058)	0.744
m21	0.113*** (0.013)	-	0.314*** (0.015)	-
m21*EB Status	-0.436*** (0.054)	1.271	-0.620*** (0.055)	1.254
m22	0.033* (0.013)	-	-3.015*** (0.117)	-
m22*EB Status	-0.734*** (0.053)	0.967	2.204*** (0.230)	0.969
m23	-0.95*** (0.013)	-	-6.073*** (0.195)	-
m23*EB Status	-0.261*** (0.052)	1.449	4.662*** (0.379)	1.477
m24	0.778*** (0.013)	-	0.958*** (0.015)	-
m24*EB Status	-1.611*** (0.053)	0.071	-1.760*** (0.054)	0.090
m25	-0.943*** (0.013)	-	-1.150*** (0.015)	-
m25*EB Status	-0.073 (0.052)	1.641	0.143** (0.054)	1.637
m26	-0.083*** (0.013)	-	0.123*** (0.015)	-
m26*EB Status	-0.533*** (0.053)	1.172	-0.714*** (0.054)	1.158
m27	-0.417*** (0.013)	-	-1.628*** (0.048)	-
m27*EB Status	0.001 (0.054)	1.717	1.173*** (0.104)	1.705
m28	0.573*** (0.013)	-	0.356*** (0.015)	-
m28*EB Status	-0.881*** (0.054)	0.816	-0.661*** (0.055)	0.816
m29	0.178*** (0.013)	-	-9.064*** (0.353)	-
m29*EB Status	0.693*** (0.062)	2.423	9.635*** (0.683)	2.501

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m30	-5.027*** (0.012)	-	-6.110*** (0.034)	-
m30*EB Status	3.939*** (0.053)	5.736	5.024*** (0.080)	6.017
m31	-0.983*** (0.013)	-	-7.408*** (0.244)	-
m31*EB Status	0.499*** (0.053)	2.225	6.681*** (0.474)	2.272
m32	0.792*** (0.013)	-	0.568*** (0.015)	-
m32*EB Status	-0.290*** (0.058)	1.420	-0.086 (0.059)	1.403
m33	-4.203*** (0.012)	-	-19.936*** (0.599)	-
m33*EB Status	1.474*** (0.053)	3.220	16.538*** (1.157)	3.254
m34	0.318*** (0.013)	-	-0.398*** (0.030)	-
m34*EB Status	-0.754*** (0.054)	0.946	-0.054 (0.074)	0.941
m35	-4.682*** (0.012)	-	-39.264*** (1.324)	-
m35*EB Status	4.821*** (0.057)	6.636	38.101*** (2.556)	6.732
m36	-0.419*** (0.013)	-	-11.332*** (0.415)	-
m36*EB Status	-0.034 (0.054)	1.681	10.462*** (0.803)	1.747
m37	-1.424*** (0.013)	-	-4.938*** (0.133)	-
m37*EB Status	1.018*** (0.054)	2.755	4.406*** (0.261)	2.799
m38	-0.434*** (0.013)	-	-25.851*** (0.97)	-
m38*EB Status	-0.475*** (0.052)	1.231	23.925*** (1.872)	1.306
m39	0.389*** (0.013)	-	0.584*** (0.015)	-
m39*EB Status	-0.430*** (0.055)	1.277	-0.615*** (0.056)	1.259
m40	-3.166*** (0.012)	-	-31.453*** (1.085)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m40*EB Status	1.567*** (0.052)	3.315	28.643*** (2.095)	3.173
m41	0.293*** (0.013)	-	-0.018 (0.017)	-
m41*EB Status	-0.090 (0.056)	1.624	0.204*** (0.060)	1.600
m42	-0.502*** (0.013)	-	-2.457*** (0.075)	-
m42*EB Status	-0.065 (0.053)	1.649	1.824*** (0.153)	1.648
delta1	4.563*** (0.004)	-	4.645*** (0.004)	-
delta1*EB Status	-1.029*** (0.017)	-	-1.035*** (0.018)	-
delta2	6.871*** (0.006)	-	7.132*** (0.006)	-
delta2*EB Status	-1.207*** (0.040)	-	-1.296*** (0.041)	-
delta3	8.092*** (0.007)	-	8.444*** (0.008)	-
delta3*EB Status	-1.058*** (0.075)	-	-1.226*** (0.075)	-
Intercept Variance	1.724		1.747	
NP Variance	-		0.112	
Intercept*Feature Covariance	-		0.385	

Note: + *p* < .10, * *p* < .05, ** *p* < .01, *** *p* < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G2.

EPvEB Models' Adjusted DIF Estimates – Mathematics Assessment

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m01*EB Status	1.086* (0.052)	[0.984, 1.188]	1.092* (0.052)	[0.990, 1.194]
m02*EB Status	1.894* (0.058)	[1.780, 2.008]	1.862* (0.057)	[1.750, 1.973]
m03*EB Status	1.494* (0.054)	[1.388, 1.600]	1.487* (0.152)	[1.189, 1.785]
m04*EB Status	1.834* (0.059)	[1.718, 1.950]	1.800* (0.058)	[1.686, 1.913]
m05*EB Status	2.157* (0.062)	[2.035, 2.279]	2.181* (0.431)	[1.336, 3.026]
m06*EB Status	1.216* (0.052)	[1.114, 1.318]	1.218* (0.052)	[1.116, 1.320]
m07*EB Status	1.584 (0.067)	[1.453, 1.715]	1.565 (0.067)	[1.433, 1.696]
m08*EB Status	1.519* (0.056)	[1.409, 1.629]	1.496* (0.057)	[1.384, 1.607]
m09*EB Status	4.580* (0.054)	[4.474, 4.686]	4.762* (0.055)	[4.654, 4.869]
m10*EB Status	1.252* (0.054)	[1.146, 1.358]	1.278* (0.616)	[0.070, 2.485]
m11*EB Status	1.182* (0.054)	[1.076, 1.288]	1.169* (0.055)	[1.061, 1.276]
m12*EB Status	3.440* (0.054)	[3.334, 3.546]	3.513* (0.171)	[3.178, 3.848]
m13*EB Status	-	-	-	-
m14*EB Status	5.628* (0.053)	[5.524, 5.732]	5.757* (0.989)	[3.818, 7.695]
m15*EB Status	1.168* (0.057)	[1.056, 1.280]	1.172* (0.474)	[0.243, 2.101]
m16*EB Status	3.165* (0.052)	[3.063, 3.267]	3.255* (0.327)	[2.614, 3.896]
m17*EB Status	1.417* (0.056)	[1.307, 1.527]	1.397* (0.055)	[1.289, 1.504]

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m18*EB Status	2.348* (0.069)	[2.213, 2.483]	2.298* (0.067)	[2.166, 2.429]
m19*EB Status	2.703* (0.053)	[2.599, 2.807]	2.756* (0.327)	[2.115, 3.397]
m20*EB Status	0.726* (0.054)	[0.620, 0.832]	0.729* (0.058)	[0.615, 0.842]
m21*EB Status	1.245* (0.054)	[1.139, 1.351]	1.229* (0.055)	[1.121, 1.336]
m22*EB Status	0.947* (0.053)	[0.843, 1.051]	0.949* (0.230)	[0.498, 1.400]
m23*EB Status	1.420* (0.052)	[1.318, 1.522]	1.447* (0.379)	[0.704, 2.190]
m24*EB Status	0.070* (0.053)	[-0.034, 0.174]	0.089* (0.054)	[-0.017, 0.194]
m25*EB Status	1.608 (0.052)	[1.506, 1.710]	1.604 (0.054)	[1.498, 1.709]
m26*EB Status	1.148* (0.053)	[1.044, 1.252]	1.135* (0.054)	[1.029, 1.240]
m27*EB Status	1.682 (0.054)	[1.576, 1.788]	1.671 (0.104)	[1.467, 1.875]
m28*EB Status	0.800* (0.054)	[0.694, 0.906]	0.800* (0.055)	[0.692, 0.907]
m29*EB Status	2.374* (0.062)	[2.252, 2.496]	2.451* (0.683)	[1.112, 3.789]
m30*EB Status	5.620* (0.053)	[5.516, 5.724]	5.896* (0.080)	[5.739, 6.053]
m31*EB Status	2.180* (0.053)	[2.076, 2.284]	2.226* (0.474)	[1.297, 3.155]
m32*EB Status	1.391* (0.058)	[1.277, 1.505]	1.375* (0.059)	[1.259, 1.490]
m33*EB Status	3.155* (0.053)	[3.051, 3.259]	3.189* (1.157)	[0.921, 5.456]
m34*EB Status	0.927* (0.054)	[0.821, 1.033]	0.922* (0.074)	[0.777, 1.067]
m35*EB Status	6.502* (0.057)	[6.390, 6.614]	6.596* (2.556)	[1.586, 11.606]

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m36*EB Status	1.647 (0.054)	[1.541, 1.753]	1.712 (0.803)	[0.138, 3.286]
m37*EB Status	2.699* (0.054)	[2.593, 2.805]	2.743* (0.261)	[2.231, 3.254]
m38*EB Status	1.206* (0.052)	[1.104, 1.308]	1.280* (1.872)	[-2.389, 4.949]
m39*EB Status	1.251* (0.055)	[1.143, 1.359]	1.234* (0.056)	[1.124, 1.343]
m40*EB Status	3.248* (0.052)	[3.146, 3.350]	3.109* (2.095)	[-0.997, 7.215]
m41*EB Status	1.591 (0.056)	[1.481, 1.701]	1.568 (0.060)	[1.45, 1.685]
m42*EB Status	1.616 (0.053)	[1.512, 1.720]	1.615 (0.153)	[1.315, 1.915]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G3.

EPvSTEB Model Results – Mathematics Assessment

Effect	Base model Estimate (SE)	Base model Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size
Intercept	-0.666*** (0.010)	-	3.934*** (0.177)	-
Intercept*STEB	1.709*** (0.054)	-	-2.481*** (0.421)	-
NP	-	-	7.248*** (0.278)	-
NP*STEB	-	-	-6.538*** (0.661)	-
m01	-1.147*** (0.013)	-	-1.261*** (0.013)	-
m01*STEB	-0.685*** (0.062)	1.045	-0.560*** (0.063)	1.047
m02	0.207*** (0.013)	-	0.099*** (0.013)	-
m02*STEB	0.152* (0.069)	1.899	0.237*** (0.069)	1.860
m03	-0.073*** (0.013)	-	-2.011*** (0.075)	-
m03*STEB	-0.281*** (0.065)	1.457	1.487*** (0.188)	1.448
m04	0.281*** (0.013)	-	0.273*** (0.012)	-
m04*STEB	0.146* (0.071)	1.893	0.135+ (0.070)	1.849
m05	0.349*** (0.013)	-	-5.443*** (0.222)	-
m05*STEB	0.421*** (0.074)	2.174	5.707*** (0.534)	2.198
m06	-1.036*** (0.013)	-	-1.146*** (0.013)	-
m06*STEB	-0.518*** (0.062)	1.216	-0.396*** (0.062)	1.214
m07	1.378*** (0.013)	-	1.531*** (0.015)	-
m07*STEB	-0.235** (0.078)	1.504	-0.410*** (0.079)	1.480
m08	0.352*** (0.013)	-	0.546*** (0.015)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m08*STEB	-0.288*** (0.067)	1.450	-0.465*** (0.069)	1.424
m09	-3.848*** (0.012)	-	-3.821*** (0.015)	-
m09*STEB	2.837*** (0.064)	4.640	2.865*** (0.066)	4.823
m10	0.148*** (0.013)	-	-8.184*** (0.319)	-
m10*STEB	-0.456*** (0.065)	1.279	7.119*** (0.763)	1.304
m11	0.258*** (0.013)	-	0.454*** (0.015)	-
m11*STEB	-0.561*** (0.065)	1.172	-0.730*** (0.067)	1.154
m12	-2.151*** (0.013)	-	-4.439*** (0.086)	-
m12*STEB	1.674*** (0.065)	3.453	3.771*** (0.212)	3.525
m13	-	-	-	-
m13*STEB	-	-	-	-
m14	-4.950*** (0.012)	-	-18.498*** (0.514)	-
m14*STEB	3.945*** (0.064)	5.770	16.213*** (1.227)	5.901
m15	0.793*** (0.013)	-	-5.577*** (0.245)	-
m15*STEB	-0.547*** (0.069)	1.186	5.255*** (0.587)	1.190
m16	-3.265*** (0.012)	-	-7.751*** (0.168)	-
m16*STEB	1.320*** (0.063)	3.091	5.389*** (0.405)	3.175
m17	0.293*** (0.013)	-	0.184*** (0.013)	-
m17*STEB	-0.422*** (0.066)	1.314	-0.320*** (0.066)	1.292
m18	0.724*** (0.013)	-	0.702*** (0.012)	-
m18*STEB	0.684*** (0.083)	2.442	0.654*** (0.082)	2.379
m19	-1.583*** (0.013)	-	-6.017*** (0.168)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m19*STEB	0.840*** (0.064)	2.601	4.876*** (0.405)	2.651
m20	0.755*** (0.013)	-	0.430*** (0.017)	-
m20*STEB	-1.082*** (0.065)	0.640	-0.775*** (0.070)	0.640
m21	0.113*** (0.013)	-	0.314*** (0.015)	-
m21*STEB	-0.527*** (0.065)	1.206	-0.698*** (0.066)	1.186
m22	0.033* (0.013)	-	-3.015*** (0.117)	-
m22*STEB	-0.893*** (0.063)	0.833	1.884*** (0.285)	0.832
m23	-0.950*** (0.013)	-	-6.072*** (0.195)	-
m23*STEB	-0.295*** (0.063)	1.443	4.357*** (0.469)	1.468
m24	0.779*** (0.013)	-	0.959*** (0.015)	-
m24*STEB	-1.665*** (0.063)	0.045	-1.800*** (0.065)	0.061
m25	-0.944*** (0.013)	-	-1.150*** (0.015)	-
m25*STEB	-0.006 (0.063)	1.738	0.200** (0.065)	1.729
m26	-0.083*** (0.013)	-	0.123*** (0.015)	-
m26*STEB	-0.626*** (0.064)	1.105	-0.793*** (0.065)	1.089
m27	-0.417*** (0.013)	-	-1.628*** (0.048)	-
m27*STEB	-0.006 (0.065)	1.738	1.103*** (0.128)	1.723
m28	0.573*** (0.013)	-	0.356*** (0.015)	-
m28*STEB	-0.896*** (0.065)	0.830	-0.686*** (0.067)	0.825
m29	0.178*** (0.013)	-	-9.062*** (0.354)	-
m29*STEB	0.596*** (0.074)	2.352	9.048*** (0.847)	2.432

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m30	-5.028*** (0.012)	-	-6.112*** (0.034)	-
m30*STEB	3.964*** (0.064)	5.790	5.007*** (0.098)	6.068
m31	-0.983*** (0.013)	-	-7.407*** (0.245)	-
m31*STEB	0.510*** (0.065)	2.265	6.352*** (0.587)	2.309
m32	0.793*** (0.013)	-	0.568*** (0.015)	-
m32*STEB	-0.394*** (0.070)	1.342	-0.199** (0.071)	1.322
m33	-4.204*** (0.012)	-	-19.935*** (0.601)	-
m33*STEB	1.399*** (0.064)	3.172	15.627*** (1.434)	3.208
m34	0.318*** (0.013)	-	-0.398*** (0.030)	-
m34*STEB	-0.826*** (0.064)	0.901	-0.161+ (0.091)	0.893
m35	-4.683*** (0.012)	-	-39.259*** (1.329)	-
m35*STEB	4.723*** (0.068)	6.564	36.128*** (3.170)	6.642
m36	-0.420*** (0.013)	-	-11.331*** (0.417)	-
m36*STEB	-0.083 (0.064)	1.659	9.832*** (0.996)	1.724
m37	-1.425*** (0.013)	-	-4.938*** (0.134)	-
m37*STEB	0.932*** (0.064)	2.695	4.135*** (0.323)	2.736
m38	-0.435*** (0.013)	-	-25.846*** (0.974)	-
m38*STEB	-0.541*** (0.063)	1.192	22.498*** (2.321)	1.265
m39	0.390*** (0.013)	-	0.584*** (0.015)	-
m39*STEB	-0.492*** (0.066)	1.242	-0.664*** (0.068)	1.221
m40	-3.167*** (0.012)	-	-31.448*** (1.089)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m40*STEB	1.501*** (0.062)	3.276	27.060*** (2.597)	3.139
m41	0.293*** (0.013)	-	-0.018 (0.017)	-
m41*STEB	-0.221** (0.067)	1.519	0.060 (0.072)	1.493
m42	-0.502*** (0.013)	-	-2.457*** (0.076)	-
m42*STEB	-0.060 (0.064)	1.683	1.726*** (0.189)	1.678
delta1	4.565*** (0.004)	-	4.646*** (0.004)	-
delta1*STEB	-1.011*** (0.020)	-	-1.013*** (0.021)	-
delta2	6.873*** (0.006)	-	7.134*** (0.006)	-
delta2*STEB	-1.236*** (0.045)	-	-1.312*** (0.047)	-
delta3	8.094*** (0.007)	-	8.447*** (0.008)	-
delta3*STEB	-1.172*** (0.082)	-	-1.325*** (0.082)	-
Intercept Variance	1.744		1.768	
NP Variance	-		0.113	
Intercept*Feature Covariance	-		0.390	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G4.

EPvSTEB Models' Adjusted DIF Estimates – Mathematics Assessment

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m01*STEB	1.024* (0.062)	[0.902, 1.146]	1.026* (0.063)	[0.902, 1.149]
m02*STEB	1.861* (0.069)	[1.726, 1.996]	1.823* (0.069)	[1.687, 1.958]
m03*STEB	1.428* (0.065)	[1.301, 1.555]	1.419* (0.188)	[1.050, 1.787]
m04*STEB	1.855* (0.071)	[1.716, 1.994]	1.812* (0.070)	[1.675, 1.949]
m05*STEB	2.130* (0.074)	[1.985, 2.275]	2.154* (0.534)	[1.107, 3.200]
m06*STEB	1.191* (0.062)	[1.069, 1.313]	1.190* (0.062)	[1.068, 1.311]
m07*STEB	1.474* (0.078)	[1.321, 1.627]	1.450* (0.079)	[1.295, 1.605]
m08*STEB	1.421* (0.067)	[1.290, 1.552]	1.395* (0.069)	[1.260, 1.530]
m09*STEB	4.546* (0.064)	[4.421, 4.671]	4.725* (0.066)	[4.596, 4.855]
m10*STEB	1.253* (0.065)	[1.126, 1.380]	1.277* (0.763)	[-0.218, 2.773]
m11*STEB	1.148* (0.065)	[1.021, 1.275]	1.130* (0.067)	[0.999, 1.262]
m12*STEB	3.383* (0.065)	[3.256, 3.510]	3.454* (0.212)	[3.039, 3.870]
m13*STEB	-	-	-	-
m14*STEB	5.654* (0.064)	[5.529, 5.779]	5.782* (1.227)	[3.377, 8.187]
m15*STEB	1.162* (0.069)	[1.027, 1.297]	1.166* (0.587)	[0.015, 2.316]
m16*STEB	3.029* (0.063)	[2.906, 3.152]	3.111* (0.405)	[2.317, 3.904]
m17*STEB	1.287* (0.066)	[1.158, 1.416]	1.266* (0.066)	[1.136, 1.395]

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m18*STEB	2.393* (0.083)	[2.230, 2.556]	2.331* (0.082)	[2.170, 2.492]
m19*STEB	2.549* (0.064)	[2.424, 2.674]	2.598* (0.405)	[1.804, 3.391]
m20*STEB	0.627* (0.065)	[0.500, 0.754]	0.628* (0.070)	[0.490, 0.765]
m21*STEB	1.182* (0.065)	[1.055, 1.309]	1.162* (0.066)	[1.033, 1.292]
m22*STEB	0.816* (0.063)	[0.693, 0.939]	0.815* (0.285)	[0.257, 1.374]
m23*STEB	1.414* (0.063)	[1.291, 1.537]	1.438* (0.469)	[0.519, 2.357]
m24*STEB	0.044* (0.063)	[-0.079, 0.167]	0.060* (0.065)	[-0.067, 0.188]
m25*STEB	1.703 (0.063)	[1.580, 1.826]	1.694 (0.065)	[1.567, 1.822]
m26*STEB	1.083* (0.064)	[0.958, 1.208]	1.067* (0.065)	[0.940, 1.195]
m27*STEB	1.703 (0.065)	[1.576, 1.830]	1.688 (0.128)	[1.437, 1.939]
m28*STEB	0.813* (0.065)	[0.686, 0.940]	0.808* (0.067)	[0.677, 0.939]
m29*STEB	2.305* (0.074)	[2.160, 2.450]	2.383* (0.847)	[0.723, 4.043]
m30*STEB	5.673* (0.064)	[5.548, 5.798]	5.945* (0.098)	[5.753, 6.137]
m31*STEB	2.219* (0.065)	[2.092, 2.346]	2.263* (0.587)	[1.112, 3.413]
m32*STEB	1.315* (0.070)	[1.178, 1.452]	1.295* (0.071)	[1.156, 1.434]
m33*STEB	3.108* (0.064)	[2.983, 3.233]	3.143* (1.434)	[0.332, 5.954]
m34*STEB	0.883* (0.064)	[0.758, 1.008]	0.875* (0.091)	[0.697, 1.054]
m35*STEB	6.432* (0.068)	[6.299, 6.565]	6.508* (3.170)	[0.295, 12.721]
m36*STEB	1.626 (0.064)	[1.501, 1.751]	1.689 (0.996)	[-0.263, 3.641]

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
m37*STEB	2.641* (0.064)	[2.516, 2.766]	2.680* (0.323)	[2.047, 3.314]
m38*STEB	1.168* (0.063)	[1.045, 1.291]	1.240* (2.321)	[-3.309, 5.789]
m39*STEB	1.217* (0.066)	[1.088, 1.346]	1.196* (0.068)	[1.063, 1.330]
m40*STEB	3.210* (0.062)	[3.088, 3.332]	3.076* (2.597)	[-2.015, 8.166]
m41*STEB	1.488* (0.067)	[1.357, 1.619]	1.463* (0.072)	[1.321, 1.604]
m42*STEB	1.649 (0.064)	[1.524, 1.774]	1.644 (0.189)	[1.274, 2.015]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G5.

EPvLTEB Model Results – Mathematics Assessment

Effect	Base model Estimate (SE)	Base model Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size
Intercept	-0.666*** (0.010)	-	3.933*** (0.177)	-
Intercept*LTEB	1.624*** (0.078)	-	-3.136*** (0.489)	-
NP	-	-	7.247*** (0.278)	-
NP*LTEB	-	-	-7.455*** (0.764)	-
m01	-1.148*** (0.013)	-	-1.261*** (0.013)	-
m01*LTEB	-0.402*** (0.090)	1.247	-0.268** (0.090)	1.258
m02	0.207*** (0.013)	-	0.099*** (0.013)	-
m02*LTEB	0.347*** (0.102)	2.012	0.444*** (0.101)	1.985
m03	-0.073*** (0.013)	-	-2.011*** (0.075)	-
m03*LTEB	0.017 (0.095)	1.675	2.022*** (0.225)	1.671
m04	0.281*** (0.013)	-	0.273*** (0.012)	-
m04*LTEB	0.171+ (0.101)	1.832	0.166+ (0.099)	1.808
m05	0.350*** (0.013)	-	-5.442*** (0.223)	-
m05*LTEB	0.596*** (0.109)	2.266	6.599*** (0.621)	2.287
m06	-1.036*** (0.013)	-	-1.146*** (0.013)	-
m06*LTEB	-0.354*** (0.090)	1.296	-0.224* (0.090)	1.303
m07	1.378*** (0.013)	-	1.531*** (0.015)	-
m07*LTEB	0.240+ (0.125)	1.902	0.040 (0.125)	1.892
m08	0.352*** (0.013)	-	0.546*** (0.015)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m08*LTEB	0.124 (0.101)	1.784	-0.088 (0.101)	1.762
m09	-3.849*** (0.012)	-	-3.821*** (0.015)	-
m09*LTEB	3.036*** (0.092)	4.756	3.026*** (0.094)	4.940
m10	0.148*** (0.013)	-	-8.183*** (0.320)	-
m10*LTEB	-0.371*** (0.094)	1.279	8.245*** (0.883)	1.303
m11	0.258*** (0.013)	-	0.454*** (0.015)	-
m11*LTEB	-0.365*** (0.094)	1.285	-0.565*** (0.096)	1.275
m12	-2.152*** (0.013)	-	-4.439*** (0.086)	-
m12*LTEB	1.945*** (0.094)	3.642	4.314*** (0.251)	3.721
m13	-	-	-	-
m13*LTEB	-	-	-	-
m14	-4.951*** (0.012)	-	-18.497*** (0.515)	-
m14*LTEB	3.954*** (0.091)	5.693	17.908*** (1.417)	5.824
m15	0.793*** (0.013)	-	-5.576*** (0.246)	-
m15*LTEB	-0.443*** (0.099)	1.205	6.153*** (0.681)	1.207
m16	-3.266*** (0.012)	-	-7.751*** (0.169)	-
m16*LTEB	1.847*** (0.090)	3.542	6.476*** (0.471)	3.645
m17	0.293*** (0.013)	-	0.184*** (0.013)	-
m17*LTEB	0.100 (0.100)	1.759	0.203* (0.099)	1.739
m18	0.725*** (0.013)	-	0.702*** (0.012)	-
m18*LTEB	0.629*** (0.118)	2.299	0.617*** (0.117)	2.268
m19	-1.584*** (0.013)	-	-6.017*** (0.169)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m19*LTEB	1.436*** (0.094)	3.123	6.022*** (0.471)	3.181
m20	0.755*** (0.013)	-	0.430*** (0.017)	-
m20*LTEB	-0.672*** (0.096)	0.972	-0.337*** (0.100)	0.975
m21	0.113*** (0.013)	-	0.314*** (0.015)	-
m21*LTEB	-0.236* (0.094)	1.417	-0.440*** (0.096)	1.402
m22	0.033** (0.013)	-	-3.014*** (0.117)	-
m22*LTEB	-0.378*** (0.093)	1.272	2.777*** (0.334)	1.277
m23	-0.950*** (0.013)	-	-6.072*** (0.196)	-
m23*LTEB	-0.190* (0.090)	1.464	5.100*** (0.544)	1.495
m24	0.779*** (0.013)	-	0.959*** (0.015)	-
m24*LTEB	-1.501*** (0.091)	0.126	-1.669*** (0.093)	0.148
m25	-0.944*** (0.013)	-	-1.150*** (0.015)	-
m25*LTEB	-0.212* (0.090)	1.441	0.017 (0.092)	1.443
m26	-0.083*** (0.013)	-	0.123*** (0.015)	-
m26*LTEB	-0.331*** (0.092)	1.320	-0.533*** (0.094)	1.307
m27	-0.417*** (0.013)	-	-1.627*** (0.048)	-
m27*LTEB	0.015 (0.093)	1.673	1.272*** (0.157)	1.666
m28	0.574*** (0.013)	-	0.356*** (0.015)	-
m28*LTEB	-0.852*** (0.093)	0.788	-0.619*** (0.095)	0.794
m29	0.178*** (0.013)	-	-9.060*** (0.355)	-
m29*LTEB	0.914*** (0.111)	2.590	10.513*** (0.980)	2.659

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m30	-5.029*** (0.012)	-	-6.113*** (0.034)	-
m30*LTEB	3.897*** (0.091)	5.635	5.038*** (0.124)	5.920
m31	-0.983*** (0.013)	-	-7.406*** (0.246)	-
m31*LTEB	0.478*** (0.092)	2.145	7.117*** (0.680)	2.191
m32	0.793*** (0.013)	-	0.568*** (0.015)	-
m32*LTEB	-0.058 (0.104)	1.598	0.158 (0.105)	1.587
m33	-4.205*** (0.012)	-	-19.933*** (0.603)	-
m33*LTEB	1.635*** (0.092)	3.326	17.828*** (1.657)	3.354
m34	0.318*** (0.013)	-	-0.398*** (0.030)	-
m34*LTEB	-0.602*** (0.093)	1.043	0.146 (0.119)	1.042
m35	-4.684*** (0.012)	-	-39.254*** (1.332)	-
m35*LTEB	5.036*** (0.101)	6.797	40.873*** (3.661)	6.931
m36	-0.420*** (0.013)	-	-11.329*** (0.418)	-
m36*LTEB	0.071 (0.093)	1.730	11.35*** (1.151)	1.794
m37	-1.425*** (0.013)	-	-4.938*** (0.134)	-
m37*LTEB	1.208*** (0.094)	2.890	4.843*** (0.377)	2.937
m38	-0.435*** (0.013)	-	-25.843*** (0.976)	-
m38*LTEB	-0.335*** (0.091)	1.316	25.909*** (2.681)	1.390
m39	0.390*** (0.013)	-	0.584*** (0.015)	-
m39*LTEB	-0.296** (0.096)	1.355	-0.500*** (0.097)	1.341
m40	-3.167*** (0.012)	-	-31.444*** (1.092)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m40*LTEB	1.709*** (0.090)	3.402	30.834*** (3.000)	3.244
m41	0.293*** (0.013)	-	-0.018 (0.017)	-
m41*LTEB	0.207* (0.101)	1.869	0.518*** (0.105)	1.848
m42	-0.502*** (0.013)	-	-2.456*** (0.076)	-
m42*LTEB	-0.075 (0.092)	1.581	1.952*** (0.225)	1.584
delta1	4.565*** (0.004)	-	4.647*** (0.004)	-
delta1*LTEB	-1.065*** (0.030)	-	-1.081*** (0.031)	-
delta2	6.874*** (0.006)	-	7.136*** (0.006)	-
delta2*LTEB	-1.092*** (0.081)	-	-1.222*** (0.083)	-
delta3	8.096*** (0.007)	-	8.449*** (0.008)	-
delta3*LTEB	-0.574** (0.189)	-	-0.791*** (0.189)	-
Intercept Variance	1.759		1.782	
NP Variance	-		0.114	
Intercept*Feature Covariance	-		0.393	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G6.

EPvLTEB Models' Adjusted DIF Estimates – Mathematics Assessment

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m01*LTEB	1.222* (0.090)	[1.046, 1.398]	1.233* (0.090)	[1.057, 1.409]
m02*LTEB	1.971* (0.102)	[1.771, 2.171]	1.945* (0.101)	[1.747, 2.143]
m03*LTEB	1.641 (0.095)	[1.455, 1.827]	1.637 (0.225)	[1.196, 2.078]
m04*LTEB	1.795 (0.101)	[1.597, 1.993]	1.771 (0.099)	[1.577, 1.965]
m05*LTEB	2.220* (0.109)	[2.006, 2.434]	2.240* (0.621)	[1.023, 3.458]
m06*LTEB	1.270* (0.090)	[1.094, 1.446]	1.277* (0.090)	[1.101, 1.453]
m07*LTEB	1.864 (0.125)	[1.619, 2.109]	1.854 (0.125)	[1.609, 2.099]
m08*LTEB	1.748 (0.101)	[1.550, 1.946]	1.726 (0.101)	[1.528, 1.924]
m09*LTEB	4.660* (0.092)	[4.480, 4.840]	4.840* (0.094)	[4.656, 5.024]
m10*LTEB	1.253* (0.094)	[1.069, 1.437]	1.277* (0.883)	[-0.454, 3.008]
m11*LTEB	1.259* (0.094)	[1.075, 1.443]	1.249* (0.096)	[1.061, 1.437]
m12*LTEB	3.569* (0.094)	[3.385, 3.753]	3.646* (0.251)	[3.154, 4.138]
m13*LTEB	-	-	-	-
m14*LTEB	5.578* (0.091)	[5.400, 5.756]	5.707* (1.417)	[2.929, 8.484]
m15*LTEB	1.181* (0.099)	[0.987, 1.375]	1.183* (0.681)	[-0.152, 2.518]
m16*LTEB	3.471* (0.090)	[3.295, 3.647]	3.571* (0.471)	[2.648, 4.494]
m17*LTEB	1.724 (0.100)	[1.528, 1.920]	1.704 (0.099)	[1.510, 1.898]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m18*LTEB	2.253* (0.118)	[2.022, 2.484]	2.222* (0.117)	[1.993, 2.452]
m19*LTEB	3.060* (0.094)	[2.876, 3.244]	3.117* (0.471)	[2.194, 4.040]
m20*LTEB	0.952* (0.096)	[0.764, 1.140]	0.955* (0.100)	[0.759, 1.151]
m21*LTEB	1.388* (0.094)	[1.204, 1.572]	1.374* (0.096)	[1.186, 1.562]
m22*LTEB	1.246* (0.093)	[1.064, 1.428]	1.251* (0.334)	[0.597, 1.906]
m23*LTEB	1.434* (0.090)	[1.258, 1.610]	1.465* (0.544)	[0.398, 2.531]
m24*LTEB	0.123* (0.091)	[-0.055, 0.301]	0.145* (0.093)	[-0.037, 0.327]
m25*LTEB	1.412* (0.090)	[1.236, 1.588]	1.414* (0.092)	[1.233, 1.594]
m26*LTEB	1.293* (0.092)	[1.113, 1.473]	1.281* (0.094)	[1.097, 1.465]
m27*LTEB	1.639 (0.093)	[1.457, 1.821]	1.632 (0.157)	[1.325, 1.940]
m28*LTEB	0.772* (0.093)	[0.590, 0.954]	0.778* (0.095)	[0.591, 0.964]
m29*LTEB	2.538* (0.111)	[2.320, 2.756]	2.606* (0.980)	[0.685, 4.527]
m30*LTEB	5.521* (0.091)	[5.343, 5.699]	5.801* (0.124)	[5.558, 6.044]
m31*LTEB	2.102* (0.092)	[1.922, 2.282]	2.147* (0.680)	[0.814, 3.480]
m32*LTEB	1.566 (0.104)	[1.362, 1.770]	1.555 (0.105)	[1.349, 1.760]
m33*LTEB	3.259* (0.092)	[3.079, 3.439]	3.286* (1.657)	[0.038, 6.534]
m34*LTEB	1.022* (0.093)	[0.840, 1.204]	1.021* (0.119)	[0.788, 1.254]
m35*LTEB	6.660* (0.101)	[6.462, 6.858]	6.791* (3.661)	[-0.384, 13.967]

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m36*LTEB	1.695 (0.093)	[1.513, 1.877]	1.758 (1.151)	[-0.498, 4.014]
m37*LTEB	2.832* (0.094)	[2.648, 3.016]	2.877* (0.377)	[2.139, 3.616]
m38*LTEB	1.289* (0.091)	[1.111, 1.467]	1.362* (2.681)	[-3.893, 6.617]
m39*LTEB	1.328* (0.096)	[1.140, 1.516]	1.314* (0.097)	[1.124, 1.504]
m40*LTEB	3.333* (0.090)	[3.157, 3.509]	3.179* (3.000)	[-2.701, 9.059]
m41*LTEB	1.831* (0.101)	[1.633, 2.029]	1.810* (0.105)	[1.604, 2.016]
m42*LTEB	1.549 (0.092)	[1.369, 1.729]	1.552 (0.225)	[1.111, 1.993]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G7.

STEBvLTEB Model Results – Mathematics Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
Intercept	1.023*** (0.049)	-	3.562*** (0.340)	-	1.661*** (0.295)	-	2.078*** (0.148)	-	3.275*** (0.515)	-
Intercept*LTEB	-0.081 (0.086)	-	-2.024*** (0.568)	-	-0.751 (0.472)	-	-0.801*** (0.236)	-	-1.351 (1.185)	-
LEX	-	-	2.429*** (0.313)	-	-	-	-	-	4.147*** (0.818)	-
LEX*LTEB	-	-	-1.824*** (0.520)	-	-	-	-	-	-0.912 (1.814)	-
NP	-	-	-	-	1.062* (0.461)	-	-	-	-3.856*** (0.722)	-
NP*LTEB	-	-	-	-	-1.074 (0.735)	-	-	-	-0.203 (1.500)	-
RC	-	-	-	-	-	-	2.587*** (0.331)	-	0.693 (0.738)	-
RC*LTEB	-	-	-	-	-	-	-1.742*** (0.520)	-	-0.455 (1.704)	-
m01	-1.797*** (0.061)	-	-1.838*** (0.061)	-	-1.794*** (0.061)	-	-1.788*** (0.060)	-	-1.860*** (0.066)	-
m01*LTEB	0.269* (0.107)	0.192	0.332** (0.107)	0.014	0.283** (0.107)	0.204	0.269* (0.107)	0.326	0.302* (0.123)	0.142
m02	0.350*** (0.068)	-	-0.246* (0.100)	-	0.325*** (0.067)	-	0.340*** (0.067)	-	-0.613** (0.205)	-
m02*LTEB	0.195 (0.121)	0.116	0.632*** (0.172)	0.092	0.209+ (0.120)	0.129	0.194 (0.119)	0.250	0.415 (0.449)	0.065

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m03	-0.346*** (0.063)	-	-1.307*** (0.142)	-	-0.613*** (0.138)	-	-0.340*** (0.062)	-	-0.991** (0.303)	-
m03*LTEB	0.291* (0.113)	0.214	1.025*** (0.240)	0.003	0.571* (0.226)	0.221	0.286* (0.111)	0.344	0.705 (0.717)	0.152
m04	0.414*** (0.069)	-	-2.735*** (0.427)	-	0.399*** (0.068)	-	0.402*** (0.068)	-	-5.091*** (1.104)	-
m04*LTEB	0.030 (0.121)	0.052	2.455*** (0.710)	0.286	0.034 (0.119)	0.035	0.033 (0.120)	0.086	1.237 (2.447)	0.121
m05	0.750*** (0.072)	-	-0.382* (0.161)	-	-0.048 (0.376)	-	0.733*** (0.071)	-	1.888*** (0.522)	-
m05*LTEB	0.179 (0.129)	0.100	1.026*** (0.273)	0.106	1.024+ (0.602)	0.099	0.181 (0.128)	0.237	0.765 (1.201)	0.051
m06	-1.524*** (0.061)	-	-1.562*** (0.060)	-	-1.520*** (0.060)	-	-1.512*** (0.060)	-	-1.582*** (0.065)	-
m06*LTEB	0.155 (0.107)	0.076	0.213* (0.107)	0.136	0.167 (0.107)	0.086	0.152 (0.107)	0.207	0.183 (0.123)	0.020
m07	1.116*** (0.076)	-	0.866*** (0.077)	-	1.110*** (0.076)	-	1.090*** (0.075)	-	0.638*** (0.103)	-
m07*LTEB	0.477** (0.145)	0.404	0.612*** (0.146)	0.203	0.447** (0.145)	0.418	0.477*** (0.144)	0.539	0.540* (0.209)	0.359
m08	0.062 (0.065)	-	-0.558*** (0.103)	-	0.089 (0.066)	-	0.061 (0.064)	-	-1.118*** (0.233)	-
m08*LTEB	0.406*** (0.119)	0.332	0.863*** (0.178)	0.111	0.368** (0.119)	0.337	0.400*** (0.117)	0.460	0.622 (0.504)	0.268
m09	-0.982*** (0.063)	-	-7.740*** (0.880)	-	-0.918*** (0.063)	-	-0.950*** (0.062)	-	-12.692*** (2.306)	-
m09*LTEB	0.187+ (0.110)	0.108	5.318*** (1.464)	0.080	0.147 (0.111)	0.111	0.175 (0.110)	0.230	2.750 (5.108)	0.071

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
m10	-0.301*** (0.063)	-	-5.599*** (0.704)	-	-1.473** (0.534)	-	-3.180*** (0.384)	-	-5.810*** (0.934)	-
m10*LTEB	0.082 (0.112)	0.001	4.147*** (1.172)	0.229	1.306 (0.852)	0.003	2.057*** (0.604)	0.123	2.852 (2.077)	0.075
m11	-0.297*** (0.063)	-	-0.945*** (0.107)	-	-0.260*** (0.064)	-	-0.291*** (0.063)	-	-1.540*** (0.247)	-
m11*LTEB	0.192+ (0.112)	0.113	0.690*** (0.182)	0.101	0.157 (0.113)	0.122	0.188+ (0.111)	0.244	0.430 (0.536)	0.055
m12	-0.462*** (0.063)	-	-5.438*** (0.658)	-	-0.764*** (0.154)	-	-9.088*** (1.136)	-	-10.133*** (1.514)	-
m12*LTEB	0.264* (0.112)	0.187	4.069*** (1.095)	0.031	0.585* (0.250)	0.193	6.189*** (1.784)	0.284	3.763 (3.293)	0.109
m13	-	-	-	-	-	-	-	-	-	-
m13*LTEB	-	-	-	-	-	-	-	-	-	-
m14	-0.977*** (0.062)	-	-8.085*** (0.927)	-	-2.921*** (0.857)	-	-9.757*** (1.137)	-	-8.354*** (1.109)	-
m14*LTEB	0.002 (0.109)	0.081	5.371*** (1.541)	0.303	1.981 (1.365)	0.078	5.915*** (1.785)	0.005	4.577* (1.857)	0.181
m15	0.241*** (0.067)	-	-1.481*** (0.237)	-	-0.647 (0.412)	-	-2.600*** (0.384)	-	-0.078 (0.540)	-
m15*LTEB	0.104 (0.118)	0.023	1.420*** (0.396)	0.192	1.039 (0.659)	0.024	2.075*** (0.606)	0.141	1.449 (0.966)	0.041
m16	-1.907*** (0.061)	-	-3.912*** (0.270)	-	-2.537*** (0.286)	-	-1.904*** (0.061)	-	-3.038*** (0.616)	-
m16*LTEB	0.513*** (0.107)	0.441	2.054*** (0.450)	0.247	1.166* (0.457)	0.458	0.522*** (0.107)	0.585	1.414 (1.472)	0.406
m17	-0.127* (0.064)	-	-0.209*** (0.063)	-	-0.139* (0.064)	-	-0.125* (0.063)	-	-0.220** (0.068)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m17*LTEB	0.513*** (0.117)	0.441	0.554*** (0.115)	0.212	0.517*** (0.116)	0.443	0.504*** (0.116)	0.566	0.526*** (0.130)	0.370
m18	1.372*** (0.081)	-	-0.328 (0.235)	-	1.328*** (0.080)	-	1.338*** (0.080)	-	-1.571** (0.586)	-
m18*LTEB	-0.041 (0.142)	0.125	1.246** (0.394)	0.334	-0.027 (0.140)	0.097	-0.031 (0.141)	0.020	0.607 (1.295)	0.171
m19	-0.723*** (0.062)	-	-5.853*** (0.675)	-	-1.339*** (0.286)	-	-0.706*** (0.061)	-	-7.243*** (1.598)	-
m19*LTEB	0.582*** (0.111)	0.511	4.518*** (1.124)	0.323	1.227** (0.458)	0.520	0.573*** (0.110)	0.637	2.674 (3.677)	0.472
m20	-0.319*** (0.063)	-	-0.473*** (0.065)	-	-0.353*** (0.065)	-	-0.312*** (0.062)	-	-0.442*** (0.081)	-
m20*LTEB	0.402*** (0.114)	0.328	0.512*** (0.117)	0.101	0.437*** (0.117)	0.331	0.394*** (0.113)	0.454	0.456** (0.168)	0.259
m21	-0.405*** (0.063)	-	-1.097*** (0.112)	-	-0.368*** (0.063)	-	-0.398*** (0.062)	-	-1.725*** (0.262)	-
m21*LTEB	0.284* (0.112)	0.207	0.815*** (0.190)	0.009	0.249* (0.113)	0.215	0.280* (0.111)	0.338	0.539 (0.570)	0.148
m22	-0.841*** (0.061)	-	-1.878*** (0.153)	-	-1.264*** (0.203)	-	-0.826*** (0.061)	-	-1.038** (0.352)	-
m22*LTEB	0.502*** (0.110)	0.430	1.301*** (0.257)	0.208	0.947** (0.327)	0.437	0.494*** (0.109)	0.556	0.978 (0.838)	0.364
m23	-1.219*** (0.061)	-	-5.602*** (0.576)	-	-1.942*** (0.330)	-	-1.205*** (0.060)	-	-6.057*** (1.334)	-
m23*LTEB	0.097 (0.107)	0.016	3.430*** (0.959)	0.196	0.847 (0.528)	0.025	0.093 (0.107)	0.147	1.901 (3.108)	0.041
m24	-0.866*** (0.061)	-	-0.902*** (0.061)	-	-0.816*** (0.062)	-	-0.849*** (0.061)	-	-1.081*** (0.074)	-
m24*LTEB	0.157 (0.109)	0.078	0.210+ (0.107)	0.139	0.121 (0.109)	0.085	0.152 (0.108)	0.207	0.172 (0.140)	0.018

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m25	-0.931*** (0.061)	-	-1.610*** (0.111)	-	-0.945*** (0.062)	-	-0.919*** (0.061)	-	-2.022*** (0.243)	-
m25*LTEB	-0.208+ (0.108)	0.295	0.330+ (0.188)	0.504	-0.180+ (0.109)	0.284	-0.211* (0.107)	0.163	0.062 (0.539)	0.350
m26	-0.693*** (0.062)	-	-1.144*** (0.088)	-	-0.647*** (0.062)	-	-0.680*** (0.061)	-	-1.611*** (0.188)	-
m26*LTEB	0.286** (0.110)	0.209	0.641*** (0.151)	0.012	0.249* (0.111)	0.215	0.280* (0.109)	0.338	0.450 (0.404)	0.145
m27	-0.413*** (0.063)	-	-1.150*** (0.117)	-	-0.577*** (0.099)	-	-0.406*** (0.062)	-	-1.059*** (0.236)	-
m27*LTEB	0.018 (0.111)	0.064	0.590** (0.197)	0.274	0.194 (0.164)	0.054	0.016 (0.109)	0.068	0.335 (0.551)	0.118
m28	-0.315*** (0.063)	-	-5.089*** (0.635)	-	-0.334*** (0.064)	-	-0.307*** (0.063)	-	-8.490*** (1.644)	-
m28*LTEB	0.042 (0.111)	0.040	3.698*** (1.056)	0.270	0.070 (0.112)	0.029	0.040 (0.110)	0.093	1.869 (3.652)	0.108
m29	0.754*** (0.072)	-	-1.489*** (0.306)	-	-0.502 (0.593)	-	0.738*** (0.071)	-	1.760* (0.879)	-
m29*LTEB	0.319* (0.131)	0.243	2.040*** (0.511)	0.022	1.670+ (0.946)	0.236	0.320* (0.130)	0.378	1.432 (2.072)	0.180
m30	-1.032*** (0.062)	-	-5.911*** (0.635)	-	-1.114*** (0.080)	-	-9.878*** (1.138)	-	-11.353*** (1.597)	-
m30*LTEB	-0.073 (0.109)	0.157	3.609*** (1.055)	0.361	0.043 (0.136)	0.149	5.852** (1.786)	0.059	3.326 (3.517)	0.227
m31	-0.463*** (0.063)	-	-5.812*** (0.710)	-	-1.363*** (0.412)	-	-0.455*** (0.062)	-	-6.344*** (1.645)	-
m31*LTEB	-0.034 (0.110)	0.117	4.055*** (1.182)	0.358	0.903 (0.657)	0.115	-0.036 (0.109)	0.015	2.181 (3.836)	0.192

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
m32	0.390*** (0.068)	-	-0.247* (0.105)	-	0.352*** (0.068)	-	0.382*** (0.067)	-	-0.590** (0.215)	-
m32*LTEB	0.334** (0.123)	0.258	0.798*** (0.181)	0.044	0.360** (0.123)	0.267	0.331** (0.122)	0.390	0.568 (0.475)	0.202
m33	-2.759*** (0.062)	-	-7.090*** (0.565)	-	-5.024*** (1.001)	-	-2.802*** (0.062)	-	-1.869 (1.551)	-
m33*LTEB	0.222* (0.110)	0.144	3.490*** (0.940)	0.066	2.529 (1.595)	0.138	0.239* (0.110)	0.296	2.301 (3.684)	0.099
m34	-0.496*** (0.062)	-	-1.763*** (0.180)	-	-0.585*** (0.076)	-	-0.486*** (0.062)	-	-2.325*** (0.419)	-
m34*LTEB	0.218* (0.111)	0.140	1.194*** (0.302)	0.075	0.318* (0.131)	0.148	0.214+ (0.110)	0.270	0.721 (0.947)	0.082
m35	0.052 (0.067)	-	-10.789*** (1.437)	-	-4.866* (2.212)	-	-8.635*** (1.138)	-	-2.484 (2.158)	-
m35*LTEB	0.301* (0.119)	0.225	8.732*** (2.389)	0.096	5.490 (3.522)	0.287	6.350*** (1.787)	0.449	7.084+ (4.254)	0.254
m36	-0.492*** (0.062)	-	-3.19*** (0.362)	-	-2.030** (0.696)	-	-3.379*** (0.383)	-	-0.136 (0.709)	-
m36*LTEB	0.149 (0.111)	0.069	2.217*** (0.603)	0.146	1.751 (1.109)	0.071	2.128*** (0.604)	0.195	1.999 (1.295)	0.008
m37	-0.479*** (0.063)	-	-6.201*** (0.757)	-	-0.962*** (0.230)	-	-0.465*** (0.062)	-	-8.544*** (1.834)	-
m37*LTEB	0.268* (0.111)	0.191	4.657*** (1.260)	0.023	0.776* (0.369)	0.198	0.261* (0.110)	0.318	2.558 (4.175)	0.136
m38	-0.956*** (0.061)	-	-4.471*** (0.466)	-	-4.596** (1.620)	-	-0.942*** (0.061)	-	6.513** (2.224)	-
m38*LTEB	0.199+ (0.108)	0.120	2.884*** (0.775)	0.090	3.956 (2.580)	0.123	0.195+ (0.107)	0.251	2.250 (4.974)	0.066

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
m39	-0.100 (0.064)	-	-1.641*** (0.214)	-	-0.068 (0.065)	-	-0.097 (0.064)	-	-2.885*** (0.549)	-
m39*LTEB	0.192+ (0.114)	0.113	1.371*** (0.358)	0.102	0.157 (0.115)	0.122	0.188+ (0.113)	0.244	0.771 (1.208)	0.055
m40	-1.632*** (0.061)	-	-7.794*** (0.804)	-	-5.751** (1.812)	-	-4.572*** (0.383)	-	2.123 (2.128)	-
m40*LTEB	0.198+ (0.108)	0.119	4.849*** (1.337)	0.106	4.381 (2.887)	0.100	2.177*** (0.603)	0.245	3.834 (4.641)	0.055
m41	0.070 (0.065)	-	-0.728*** (0.123)	-	0.026 (0.067)	-	0.068 (0.065)	-	-1.149*** (0.269)	-
m41*LTEB	0.422*** (0.119)	0.348	1.017*** (0.209)	0.128	0.459*** (0.121)	0.353	0.416*** (0.118)	0.476	0.721 (0.601)	0.285
m42	-0.550*** (0.062)	-	-4.498*** (0.524)	-	-0.817*** (0.139)	-	-0.540*** (0.062)	-	-6.346*** (1.282)	-
m42*LTEB	-0.018 (0.110)	0.101	2.997*** (0.872)	0.325	0.267 (0.225)	0.092	-0.020 (0.108)	0.031	1.538 (2.902)	0.165
delta1	3.488*** (0.019)	-	3.623*** (0.020)	-	3.550*** (0.020)	-	3.579*** (0.020)	-	3.653*** (0.021)	-
delta1*LTEB	-0.029 (0.035)	-	-0.067+ (0.037)	-	-0.043 (0.036)	-	-0.035 (0.037)	-	-0.056 (0.038)	-
delta2	5.555*** (0.045)	-	5.856*** (0.047)	-	5.698*** (0.046)	-	5.829*** (0.047)	-	5.973*** (0.048)	-
delta2*LTEB	0.177+ (0.093)	-	0.063 (0.095)	-	0.129 (0.094)	-	0.115 (0.096)	-	0.057 (0.098)	-
delta3	6.838*** (0.082)	-	7.211*** (0.084)	-	6.991*** (0.082)	-	7.241*** (0.086)	-	7.415*** (0.087)	-
delta3*LTEB	0.632** (0.207)	-	0.475* (0.209)	-	0.576** (0.206)	-	0.519* (0.214)	-	0.437* (0.215)	-
Intercept Variance	0.791		0.835		0.799		0.831		0.846	

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
LEX Variance	-		0.118		-		-		0.058	
NP Variance	-		-		0.036		-		0.027	
RC Variance	-		-		-		0.121		0.068	
Intercept*Feature Covariance	-		0.287		0.168		0.278		See Table G15	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G8.

STEBvLTEB Models' Adjusted DIF Estimates – Mathematics Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m01*LTEB	0.188* (0.107)	[-0.022, 0.398]	-0.014* (0.107)	[-0.224, 0.196]	0.200* (0.107)	[-0.010, 0.410]	0.320* (0.107)	[0.110, 0.530]	0.139* (0.123)	[-0.102, 0.380]
m02*LTEB	0.114 (0.121)	[-0.123, 0.351]	-0.090* (0.172)	[-0.427, 0.247]	0.126* (0.120)	[-0.109, 0.361]	0.245* (0.119)	[0.012, 0.478]	0.064* (0.449)	[-0.816, 0.944]
m03*LTEB	0.210* (0.113)	[-0.011, 0.431]	-0.003* (0.240)	[-0.473, 0.467]	0.216* (0.226)	[-0.227, 0.659]	0.337* (0.111)	[0.119, 0.554]	0.149* (0.717)	[-1.256, 1.555]
m04*LTEB	-0.051 (0.121)	[-0.288, 0.186]	-0.280* (0.710)	[-1.672, 1.111]	-0.034* (0.119)	[-0.267, 0.199]	0.084* (0.120)	[-0.151, 0.319]	-0.118 (2.447)	[-4.914, 4.678]
m05*LTEB	0.098 (0.129)	[-0.155, 0.351]	-0.104* (0.273)	[-0.639, 0.431]	0.097 (0.602)	[-1.083, 1.277]	0.232* (0.128)	[-0.019, 0.483]	0.050 (1.201)	[-2.304, 2.404]
m06*LTEB	0.074 (0.107)	[-0.136, 0.284]	-0.133* (0.107)	[-0.343, 0.077]	0.084* (0.107)	[-0.126, 0.294]	0.203* (0.107)	[-0.007, 0.413]	0.020* (0.123)	[-0.221, 0.261]
m07*LTEB	0.396* (0.145)	[0.112, 0.680]	0.199* (0.146)	[-0.088, 0.485]	0.409* (0.145)	[0.125, 0.693]	0.528* (0.144)	[0.246, 0.810]	0.352* (0.209)	[-0.058, 0.761]
m08*LTEB	0.325* (0.119)	[0.092, 0.558]	0.109* (0.178)	[-0.240, 0.457]	0.330* (0.119)	[0.097, 0.563]	0.451* (0.117)	[0.222, 0.680]	0.263* (0.504)	[-0.725, 1.251]
m09*LTEB	0.106 (0.110)	[-0.110, 0.322]	-0.079 (1.464)	[-2.948, 2.791]	0.109* (0.111)	[-0.108, 0.327]	0.226* (0.110)	[0.010, 0.441]	0.070 (5.108)	[-9.942, 10.082]
m10*LTEB	0.001 (0.112)	[-0.219, 0.221]	-0.224 (1.172)	[-2.522, 2.073]	0.003 (0.852)	[-1.667, 1.673]	0.120 (0.604)	[-1.064, 1.304]	-0.074 (2.077)	[-4.145, 3.997]
m11*LTEB	0.111 (0.112)	[-0.109, 0.331]	-0.099* (0.182)	[-0.456, 0.258]	0.119* (0.113)	[-0.102, 0.341]	0.239* (0.111)	[0.021, 0.456]	0.054* (0.536)	[-0.997, 1.104]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m12*LTEB	0.183* (0.112)	[-0.037, 0.403]	-0.031 (1.095)	[-2.177, 2.115]	0.189* (0.250)	[-0.301, 0.679]	0.279 (1.784)	[-3.218, 3.775]	0.107 (3.293)	[-6.347, 6.561]
m13*LTEB	-	-	-	-	-	-	-	-	-	-
m14*LTEB	-0.079 (0.109)	[-0.293, 0.135]	-0.297 (1.541)	[-3.318, 2.723]	-0.076 (1.365)	[-2.751, 2.599]	0.005 (1.785)	[-3.494, 3.503]	-0.178 (1.857)	[-3.817, 3.462]
m15*LTEB	0.023 (0.118)	[-0.208, 0.254]	-0.188* (0.396)	[-0.964, 0.588]	0.024 (0.659)	[-1.268, 1.315]	0.138 (0.606)	[-1.050, 1.326]	-0.041 (0.966)	[-1.934, 1.853]
m16*LTEB	0.432* (0.107)	[0.222, 0.642]	0.242* (0.450)	[-0.640, 1.124]	0.448* (0.457)	[-0.447, 1.344]	0.573* (0.107)	[0.363, 0.783]	0.398 (1.472)	[-2.488, 3.283]
m17*LTEB	0.432* (0.117)	[0.203, 0.661]	0.208* (0.115)	[-0.017, 0.433]	0.434* (0.116)	[0.207, 0.661]	0.555* (0.116)	[0.327, 0.782]	0.363* (0.130)	[0.108, 0.618]
m18*LTEB	-0.122 (0.142)	[-0.400, 0.156]	-0.327* (0.394)	[-1.100, 0.445]	-0.095* (0.140)	[-0.369, 0.179]	0.020* (0.141)	[-0.257, 0.296]	-0.167 (1.295)	[-2.705, 2.371]
m19*LTEB	0.501* (0.111)	[0.283, 0.719]	0.316* (1.124)	[-1.887, 2.519]	0.509* (0.458)	[-0.388, 1.407]	0.624* (0.110)	[0.408, 0.839]	0.463 (3.677)	[-6.744, 7.670]
m20*LTEB	0.321* (0.114)	[0.098, 0.544]	0.099* (0.117)	[-0.131, 0.328]	0.324* (0.117)	[0.095, 0.553]	0.445* (0.113)	[0.223, 0.666]	0.253* (0.168)	[-0.076, 0.583]
m21*LTEB	0.203* (0.112)	[-0.017, 0.423]	-0.009* (0.190)	[-0.381, 0.364]	0.211* (0.113)	[-0.010, 0.433]	0.331* (0.111)	[0.113, 0.548]	0.145* (0.570)	[-0.972, 1.263]
m22*LTEB	0.421* (0.110)	[0.205, 0.637]	0.204* (0.257)	[-0.300, 0.707]	0.428* (0.327)	[-0.213, 1.069]	0.545* (0.109)	[0.331, 0.758]	0.357* (0.838)	[-1.286, 1.999]
m23*LTEB	0.016 (0.107)	[-0.194, 0.226]	-0.192 (0.959)	[-2.071, 1.688]	0.024 (0.528)	[-1.011, 1.059]	0.144* (0.107)	[-0.066, 0.354]	-0.040 (3.108)	[-6.132, 6.052]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
m24*LTEB	0.076 (0.109)	[-0.138, 0.290]	-0.136* (0.107)	[-0.346, 0.074]	0.083* (0.109)	[-0.131, 0.297]	0.203* (0.108)	[-0.009, 0.415]	0.017* (0.140)	[-0.257, 0.292]
m25*LTEB	-0.289 (0.108)	[-0.501, -0.077]	-0.494* (0.188)	[-0.862, -0.125]	-0.278* (0.109)	[-0.492, -0.064]	-0.160* (0.107)	[-0.370, 0.050]	-0.343 (0.539)	[-1.399, 0.713]
m26*LTEB	0.205* (0.110)	[-0.011, 0.421]	-0.011* (0.151)	[-0.307, 0.285]	0.211* (0.111)	[-0.006, 0.429]	0.331* (0.109)	[0.117, 0.544]	0.142* (0.404)	[-0.650, 0.934]
m27*LTEB	-0.063 (0.111)	[-0.281, 0.155]	-0.268* (0.197)	[-0.655, 0.118]	-0.053* (0.164)	[-0.375, 0.268]	0.067* (0.109)	[-0.147, 0.280]	-0.116* (0.551)	[-1.195, 0.964]
m28*LTEB	-0.039 (0.111)	[-0.257, 0.179]	-0.265 (1.056)	[-2.335, 1.805]	-0.028* (0.112)	[-0.248, 0.192]	0.091* (0.110)	[-0.125, 0.306]	-0.106 (3.652)	[-7.263, 7.052]
m29*LTEB	0.238* (0.131)	[-0.019, 0.495]	0.021* (0.511)	[-0.980, 1.023]	0.232 (0.946)	[-1.623, 2.086]	0.371* (0.130)	[0.116, 0.626]	0.176 (2.072)	[-3.885, 4.237]
m30*LTEB	-0.154 (0.109)	[-0.368, 0.060]	-0.354 (1.055)	[-2.422, 1.714]	-0.146* (0.136)	[-0.413, 0.120]	-0.058 (1.786)	[-3.559, 3.442]	-0.223 (3.517)	[-7.116, 6.671]
m31*LTEB	-0.115 (0.110)	[-0.331, 0.101]	-0.351 (1.182)	[-2.668, 1.966]	-0.112 (0.657)	[-1.400, 1.176]	0.015* (0.109)	[-0.199, 0.228]	-0.189 (3.836)	[-7.707, 7.330]
m32*LTEB	0.253* (0.123)	[0.012, 0.494]	0.044* (0.181)	[-0.311, 0.398]	0.262* (0.123)	[0.021, 0.503]	0.382* (0.122)	[0.143, 0.621]	0.198* (0.475)	[-0.733, 1.129]
m33*LTEB	0.141* (0.110)	[-0.075, 0.357]	-0.064* (0.940)	[-1.907, 1.778]	0.135 (1.595)	[-2.991, 3.261]	0.290* (0.110)	[0.074, 0.505]	0.097 (3.684)	[-7.124, 7.317]
m34*LTEB	0.137* (0.111)	[-0.081, 0.355]	-0.073* (0.302)	[-0.665, 0.519]	0.145* (0.131)	[-0.112, 0.402]	0.265* (0.110)	[0.049, 0.480]	0.080 (0.947)	[-1.776, 1.936]
m35*LTEB	0.220* (0.119)	[-0.013, 0.453]	0.094 (2.389)	[-4.588, 4.777]	0.281 (3.522)	[-6.622, 7.184]	0.440 (1.787)	[-3.063, 3.942]	0.249 (4.254)	[-8.089, 8.587]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
m36*LTEB	0.068 (0.111)	[-0.150, 0.286]	-0.143* (0.603)	[-1.324, 1.039]	0.070 (1.109)	[-2.104, 2.244]	0.191 (0.604)	[-0.993, 1.375]	0.008 (1.295)	[-2.530, 2.546]
m37*LTEB	0.187* (0.111)	[-0.031, 0.405]	-0.023 (1.260)	[-2.492, 2.447]	0.194* (0.369)	[-0.530, 0.917]	0.312* (0.110)	[0.096, 0.527]	0.133 (4.175)	[-8.050, 8.316]
m38*LTEB	0.118 (0.108)	[-0.094, 0.330]	-0.088* (0.775)	[-1.607, 1.431]	0.120 (2.580)	[-4.936, 5.177]	0.246* (0.107)	[0.036, 0.456]	0.064 (4.974)	[-9.685, 9.813]
m39*LTEB	0.111 (0.114)	[-0.112, 0.334]	-0.100* (0.358)	[-0.802, 0.601]	0.119* (0.115)	[-0.106, 0.345]	0.239* (0.113)	[0.017, 0.460]	0.054 (1.208)	[-2.314, 2.421]
m40*LTEB	0.117 (0.108)	[-0.095, 0.329]	-0.104 (1.337)	[-2.725, 2.516]	0.098 (2.887)	[-5.561, 5.756]	0.240 (0.603)	[-0.942, 1.422]	0.054 (4.641)	[-9.042, 9.150]
m41*LTEB	0.341* (0.119)	[0.108, 0.574]	0.126* (0.209)	[-0.284, 0.535]	0.346* (0.121)	[0.109, 0.583]	0.467* (0.118)	[0.236, 0.698]	0.279* (0.601)	[-0.899, 1.457]
m42*LTEB	-0.099 (0.110)	[-0.315, 0.117]	-0.318 (0.872)	[-2.028, 1.391]	-0.090* (0.225)	[-0.531, 0.351]	0.031* (0.108)	[-0.181, 0.243]	-0.162 (2.902)	[-5.850, 5.526]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G9.

EPvSPA Model Results – Mathematics Assessment

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	-0.666*** (0.010)	-	3.934*** (0.176)	-
Intercept*SPA	1.868*** (0.058)	-	-2.829*** (0.374)	-
NP	-	-	7.247*** (0.277)	-
NP*SPA	-	-	-7.355*** (0.586)	-
m01	-1.147*** (0.013)	-	-1.261*** (0.013)	-
m01*SPA	-0.634*** (0.067)	1.259	-0.499*** (0.067)	1.272
m02	0.207*** (0.013)	-	0.099*** (0.013)	-
m02*SPA	0.440*** (0.079)	2.356	0.535*** (0.079)	2.328
m03	-0.073*** (0.013)	-	-2.011*** (0.075)	-
m03*SPA	-0.206** (0.071)	1.696	1.776*** (0.171)	1.695
m04	0.281*** (0.013)	-	0.273*** (0.012)	-
m04*SPA	0.229** (0.078)	2.140	0.223** (0.078)	2.114
m05	0.349*** (0.013)	-	-5.443*** (0.222)	-
m05*SPA	0.574*** (0.084)	2.492	6.504*** (0.476)	2.520
m06	-1.036*** (0.013)	-	-1.145*** (0.013)	-
m06*SPA	-0.458*** (0.067)	1.439	-0.327*** (0.067)	1.448
m07	1.378*** (0.013)	-	1.530*** (0.015)	-
m07*SPA	-0.193* (0.088)	1.709	-0.377*** (0.089)	1.712
m08	0.352*** (0.013)	-	0.545*** (0.015)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m08*SPA	-0.177* (0.074)	1.726	-0.378*** (0.075)	1.711
m09	-3.848*** (0.012)	-	-3.820*** (0.015)	-
m09*SPA	3.324*** (0.071)	5.299	3.310*** (0.072)	5.475
m10	0.148*** (0.013)	-	-8.183*** (0.319)	-
m10*SPA	-0.488*** (0.070)	1.408	8.018*** (0.677)	1.438
m11	0.258*** (0.013)	-	0.454*** (0.015)	-
m11*SPA	-0.594*** (0.070)	1.300	-0.787*** (0.071)	1.294
m12	-2.151*** (0.013)	-	-4.438*** (0.085)	-
m12*SPA	1.659*** (0.070)	3.600	4.000*** (0.192)	3.680
m13	-	-	-	-
m13*SPA	-	-	-	-
m14	-4.949*** (0.012)	-	-18.497*** (0.513)	-
m14*SPA	4.214*** (0.069)	6.207	18.013*** (1.087)	6.369
m15	0.793*** (0.013)	-	-5.576*** (0.244)	-
m15*SPA	-0.672*** (0.074)	1.221	5.839*** (0.522)	1.225
m16	-3.265*** (0.012)	-	-7.750*** (0.168)	-
m16*SPA	1.676*** (0.067)	3.617	6.247*** (0.361)	3.721
m17	0.293*** (0.013)	-	0.184*** (0.013)	-
m17*SPA	-0.182* (0.073)	1.721	-0.073 (0.073)	1.707
m18	0.724*** (0.013)	-	0.702*** (0.012)	-
m18*SPA	0.728*** (0.095)	2.649	0.717*** (0.094)	2.619
m19	-1.583*** (0.013)	-	-6.017*** (0.168)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m19*SPA	0.963*** (0.069)	2.889	5.490*** (0.361)	2.948
m20	0.755*** (0.013)	-	0.430*** (0.017)	-
m20*SPA	-1.063*** (0.071)	0.822	-0.724*** (0.074)	0.833
m21	0.113*** (0.013)	-	0.314*** (0.015)	-
m21*SPA	-0.411*** (0.071)	1.487	-0.609*** (0.072)	1.475
m22	0.033* (0.013)	-	-3.015*** (0.117)	-
m22*SPA	-0.914*** (0.068)	0.974	2.203*** (0.255)	0.983
m23	-0.950*** (0.013)	-	-6.072*** (0.195)	-
m23*SPA	-0.276*** (0.067)	1.625	4.947*** (0.417)	1.659
m24	0.779*** (0.013)	-	0.958*** (0.015)	-
m24*SPA	-1.771*** (0.068)	0.099	-1.933*** (0.069)	0.124
m25	-0.943*** (0.013)	-	-1.150*** (0.015)	-
m25*SPA	-0.199** (0.068)	1.703	0.029 (0.069)	1.706
m26	-0.083*** (0.013)	-	0.123*** (0.015)	-
m26*SPA	-0.605*** (0.069)	1.289	-0.799*** (0.070)	1.282
m27	-0.417*** (0.013)	-	-1.627*** (0.048)	-
m27*SPA	0.049 (0.070)	1.956	1.290*** (0.120)	1.950
m28	0.573*** (0.013)	-	0.356*** (0.015)	-
m28*SPA	-0.928*** (0.070)	0.959	-0.697*** (0.071)	0.965
m29	0.178*** (0.013)	-	-9.061*** (0.353)	-
m29*SPA	0.675*** (0.083)	2.595	10.149*** (0.752)	2.667

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m30	-5.028*** (0.012)	-	-6.111*** (0.034)	-
m30*SPA	4.068*** (0.069)	6.058	5.195*** (0.095)	6.341
m31	-0.983*** (0.013)	-	-7.406*** (0.245)	-
m31*SPA	0.525*** (0.070)	2.442	7.082*** (0.521)	2.494
m32	0.792*** (0.013)	-	0.568*** (0.015)	-
m32*SPA	-0.240** (0.078)	1.662	-0.022 (0.079)	1.654
m33	-4.204*** (0.012)	-	-19.933*** (0.600)	-
m33*SPA	1.648*** (0.068)	3.588	17.618*** (1.271)	3.609
m34	0.318*** (0.013)	-	-0.398*** (0.030)	-
m34*SPA	-0.806*** (0.070)	1.084	-0.065 (0.090)	1.085
m35	-4.683*** (0.012)	-	-39.255*** (1.326)	-
m35*SPA	5.223*** (0.080)	7.237	40.679*** (2.808)	7.470
m36	-0.419*** (0.013)	-	-11.33*** (0.416)	-
m36*SPA	-0.047 (0.070)	1.858	11.088*** (0.883)	1.928
m37	-1.424*** (0.013)	-	-4.937*** (0.133)	-
m37*SPA	0.988*** (0.070)	2.915	4.578*** (0.289)	2.964
m38	-0.434*** (0.013)	-	-25.844*** (0.971)	-
m38*SPA	-0.438*** (0.068)	1.459	25.464*** (2.056)	1.543
m39	0.390*** (0.013)	-	0.584*** (0.015)	-
m39*SPA	-0.571*** (0.071)	1.324	-0.765*** (0.072)	1.316
m40	-3.167*** (0.012)	-	-31.445*** (1.087)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**
m40*SPA	1.648*** (0.067)	3.588	30.378*** (2.301)	3.428
m41	0.293*** (0.013)	-	-0.018 (0.017)	-
m41*SPA	-0.134+ (0.074)	1.770	0.181* (0.077)	1.756
m42	-0.502*** (0.013)	-	-2.456*** (0.076)	-
m42*SPA	-0.188** (0.069)	1.715	1.815*** (0.172)	1.720
delta1	4.564*** (0.004)	-	4.646*** (0.004)	-
delta1*SPA	-1.059*** (0.023)	-	-1.075*** (0.024)	-
delta2	6.872*** (0.006)	-	7.134*** (0.006)	-
delta2*SPA	-0.952*** (0.068)	-	-1.083*** (0.070)	-
delta3	8.093*** (0.007)	-	8.446*** (0.008)	-
delta3*SPA	-0.342* (0.167)	-	-0.561*** (0.167)	-
Intercept Variance	1.736		1.759	
NP Variance	-		0.113	
Intercept*Feature Covariance	-		0.389	

Note: + p < .10, * p < .05, ** p < .01, *** p < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G10.

EPvSPA Models' Adjusted DIF Estimates – Mathematics Assessment

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m01*SPA	1.234* (0.067)	[1.103, 1.365]	1.247* (0.067)	[1.115, 1.378]
m02*SPA	2.308* (0.079)	[2.153, 2.463]	2.281* (0.079)	[2.126, 2.436]
m03*SPA	1.662* (0.071)	[1.523, 1.801]	1.661* (0.171)	[1.326, 1.996]
m04*SPA	2.097* (0.078)	[1.944, 2.250]	2.072* (0.078)	[1.919, 2.225]
m05*SPA	2.442* (0.084)	[2.277, 2.607]	2.469* (0.476)	[1.536, 3.402]
m06*SPA	1.410* (0.067)	[1.279, 1.541]	1.419* (0.067)	[1.287, 1.550]
m07*SPA	1.675* (0.088)	[1.503, 1.847]	1.678* (0.089)	[1.503, 1.852]
m08*SPA	1.691* (0.074)	[1.546, 1.836]	1.677* (0.075)	[1.530, 1.824]
m09*SPA	5.192* (0.071)	[5.053, 5.331]	5.365* (0.072)	[5.224, 5.506]
m10*SPA	1.380* (0.070)	[1.243, 1.517]	1.409* (0.677)	[0.082, 2.735]
m11*SPA	1.274* (0.070)	[1.137, 1.411]	1.268* (0.071)	[1.129, 1.407]
m12*SPA	3.527* (0.070)	[3.390, 3.664]	3.606* (0.192)	[3.229, 3.982]
m13*SPA	-	-	-	-
m14*SPA	6.082* (0.069)	[5.947, 6.217]	6.240* (1.087)	[4.110, 8.371]
m15*SPA	1.196* (0.074)	[1.051, 1.341]	1.201* (0.522)	[0.178, 2.224]
m16*SPA	3.544* (0.067)	[3.413, 3.675]	3.646* (0.361)	[2.938, 4.354]
m17*SPA	1.686* (0.073)	[1.543, 1.829]	1.673* (0.073)	[1.530, 1.816]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m18*SPA	2.596* (0.095)	[2.410, 2.782]	2.566* (0.094)	[2.382, 2.750]
m19*SPA	2.831* (0.069)	[2.696, 2.966]	2.889* (0.361)	[2.181, 3.597]
m20*SPA	0.805* (0.071)	[0.666, 0.944]	0.816* (0.074)	[0.671, 0.961]
m21*SPA	1.457* (0.071)	[1.318, 1.596]	1.446* (0.072)	[1.305, 1.587]
m22*SPA	0.954* (0.068)	[0.821, 1.087]	0.963* (0.255)	[0.463, 1.462]
m23*SPA	1.592* (0.067)	[1.461, 1.723]	1.625* (0.417)	[0.808, 2.443]
m24*SPA	0.097* (0.068)	[-0.036, 0.230]	0.122* (0.069)	[-0.014, 0.257]
m25*SPA	1.669* (0.068)	[1.536, 1.802]	1.672* (0.069)	[1.537, 1.807]
m26*SPA	1.263* (0.069)	[1.128, 1.398]	1.256* (0.070)	[1.119, 1.393]
m27*SPA	1.917 (0.070)	[1.780, 2.054]	1.910 (0.120)	[1.675, 2.146]
m28*SPA	0.940* (0.070)	[0.803, 1.077]	0.946* (0.071)	[0.807, 1.085]
m29*SPA	2.543* (0.083)	[2.380, 2.706]	2.613* (0.752)	[1.139, 4.087]
m30*SPA	5.936* (0.069)	[5.801, 6.071]	6.213* (0.095)	[6.026, 6.399]
m31*SPA	2.393* (0.070)	[2.256, 2.530]	2.444* (0.521)	[1.423, 3.465]
m32*SPA	1.628* (0.078)	[1.475, 1.781]	1.621* (0.079)	[1.466, 1.776]
m33*SPA	3.516* (0.068)	[3.383, 3.649]	3.536* (1.271)	[1.045, 6.027]
m34*SPA	1.062* (0.070)	[0.925, 1.199]	1.063* (0.090)	[0.887, 1.239]
m35*SPA	7.091* (0.080)	[6.934, 7.248]	7.319* (2.808)	[1.816, 12.823]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m36*SPA	1.821 (0.070)	[1.684, 1.958]	1.890 (0.883)	[0.159, 3.620]
m37*SPA	2.856* (0.070)	[2.719, 2.993]	2.904* (0.289)	[2.337, 3.470]
m38*SPA	1.430* (0.068)	[1.297, 1.563]	1.511* (2.056)	[-2.518, 5.541]
m39*SPA	1.297* (0.071)	[1.158, 1.436]	1.290* (0.072)	[1.149, 1.431]
m40*SPA	3.516* (0.067)	[3.385, 3.647]	3.358* (2.301)	[-1.152, 7.868]
m41*SPA	1.734 (0.074)	[1.589, 1.879]	1.721 (0.077)	[1.570, 1.872]
m42*SPA	1.680* (0.069)	[1.545, 1.815]	1.685* (0.172)	[1.348, 2.022]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G11.

EPvOTH Model Results – Mathematics Assessment

Effect	Base model Estimate (SE)	Base model Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size
Intercept	-0.666*** (0.010)	-	3.934*** (0.177)	-
Intercept*OTH	1.404*** (0.068)	-	-2.154*** (0.674)	-
NP	-	-	7.248*** (0.279)	-
NP*OTH	-	-	-5.544*** (1.059)	-
m01	-1.148*** (0.013)	-	-1.261*** (0.013)	-
m01*OTH	-0.549*** (0.079)	0.873	-0.443*** (0.080)	0.869
m02	0.207*** (0.013)	-	0.099*** (0.013)	-
m02*OTH	-0.033 (0.084)	1.399	0.044 (0.084)	1.366
m03	-0.073*** (0.013)	-	-2.011*** (0.075)	-
m03*OTH	-0.163* (0.082)	1.267	1.335*** (0.294)	1.252
m04	0.281*** (0.013)	-	0.273*** (0.012)	-
m04*OTH	0.068 (0.087)	1.502	0.060 (0.085)	1.461
m05	0.350*** (0.013)	-	-5.443*** (0.223)	-
m05*OTH	0.379*** (0.090)	1.820	4.860*** (0.852)	1.834
m06	-1.036*** (0.013)	-	-1.146*** (0.013)	-
m06*OTH	-0.488*** (0.079)	0.935	-0.384*** (0.08)	0.929
m07	1.379*** (0.013)	-	1.531*** (0.015)	-
m07*OTH	0.022 (0.099)	1.455	-0.140 (0.101)	1.416
m08	0.352*** (0.013)	-	0.546*** (0.015)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m08*OTH	-0.141+ (0.085)	1.289	-0.296*** (0.088)	1.257
m09	-3.849*** (0.012)	-	-3.821*** (0.015)	-
m09*OTH	2.298*** (0.081)	3.778	2.352*** (0.086)	3.959
m10	0.148*** (0.013)	-	-8.183*** (0.321)	-
m10*OTH	-0.348*** (0.082)	1.078	6.078*** (1.221)	1.097
m11	0.258*** (0.013)	-	0.454*** (0.015)	-
m11*OTH	-0.366*** (0.083)	1.059	-0.513*** (0.086)	1.035
m12	-2.152*** (0.013)	-	-4.439*** (0.086)	-
m12*OTH	1.903*** (0.082)	3.375	3.691*** (0.333)	3.441
m13	-	-	-	-
m13*OTH	-	-	-	-
m14	-4.951*** (0.012)	-	-18.499*** (0.516)	-
m14*OTH	3.554*** (0.080)	5.060	13.946*** (1.964)	5.154
m15	0.794*** (0.013)	-	-5.576*** (0.246)	-
m15*OTH	-0.292*** (0.087)	1.135	4.630*** (0.938)	1.135
m16	-3.266*** (0.012)	-	-7.751*** (0.169)	-
m16*OTH	1.183*** (0.080)	2.640	4.638*** (0.646)	2.711
m17	0.293*** (0.013)	-	0.184*** (0.013)	-
m17*OTH	-0.369*** (0.083)	1.056	-0.282*** (0.083)	1.033
m18	0.725*** (0.013)	-	0.702*** (0.012)	-
m18*OTH	0.617*** (0.098)	2.063	0.586*** (0.096)	1.998
m19	-1.584*** (0.013)	-	-6.018*** (0.169)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m19*OTH	1.106*** (0.081)	2.562	4.537*** (0.646)	2.607
m20	0.755*** (0.013)	-	0.431*** (0.017)	-
m20*OTH	-0.805*** (0.083)	0.611	-0.547*** (0.093)	0.604
m21	0.113*** (0.013)	-	0.314*** (0.015)	-
m21*OTH	-0.472*** (0.082)	0.951	-0.617*** (0.085)	0.929
m22	0.033** (0.013)	-	-3.015*** (0.118)	-
m22*OTH	-0.467*** (0.081)	0.956	1.888*** (0.452)	0.951
m23	-0.950*** (0.013)	-	-6.072*** (0.196)	-
m23*OTH	-0.247** (0.079)	1.181	3.700*** (0.749)	1.199
m24	0.779*** (0.013)	-	0.959*** (0.015)	-
m24*OTH	-1.376*** (0.081)	0.029	-1.489*** (0.085)	0.039
m25	-0.944*** (0.013)	-	-1.151*** (0.015)	-
m25*OTH	0.110 (0.080)	1.545	0.285*** (0.085)	1.533
m26	-0.083*** (0.013)	-	0.123*** (0.015)	-
m26*OTH	-0.433*** (0.081)	0.991	-0.576*** (0.085)	0.971
m27	-0.417*** (0.013)	-	-1.628*** (0.048)	-
m27*OTH	-0.067 (0.081)	1.365	0.876*** (0.194)	1.349
m28	0.574*** (0.013)	-	0.356*** (0.015)	-
m28*OTH	-0.819*** (0.082)	0.597	-0.638*** (0.086)	0.591
m29	0.178*** (0.013)	-	-9.061*** (0.356)	-
m29*OTH	0.721*** (0.092)	2.169	7.900*** (1.355)	2.243

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m30	-5.029*** (0.012)	-	-6.113*** (0.034)	-
m30*OTH	3.746*** (0.081)	5.256	4.677*** (0.144)	5.534
m31	-0.983*** (0.013)	-	-7.407*** (0.246)	-
m31*OTH	0.461*** (0.081)	1.903	5.418*** (0.938)	1.939
m32	0.793*** (0.013)	-	0.568*** (0.015)	-
m32*OTH	-0.346*** (0.087)	1.080	-0.182* (0.090)	1.056
m33	-4.205*** (0.012)	-	-19.935*** (0.604)	-
m33*OTH	1.215*** (0.081)	2.673	13.297*** (2.296)	2.715
m34	0.318*** (0.013)	-	-0.398*** (0.030)	-
m34*OTH	-0.684*** (0.081)	0.735	-0.120 (0.131)	0.723
m35	-4.684*** (0.012)	-	-39.258*** (1.334)	-
m35*OTH	4.349*** (0.083)	5.871	30.837*** (5.075)	5.787
m36	-0.420*** (0.013)	-	-11.33*** (0.419)	-
m36*OTH	-0.018 (0.081)	1.415	8.395*** (1.593)	1.470
m37	-1.425*** (0.013)	-	-4.938*** (0.134)	-
m37*OTH	1.059*** (0.081)	2.514	3.783*** (0.514)	2.551
m38	-0.435*** (0.013)	-	-25.845*** (0.977)	-
m38*OTH	-0.537*** (0.080)	0.885	18.994*** (3.717)	0.937
m39	0.390*** (0.013)	-	0.584*** (0.015)	-
m39*OTH	-0.231** (0.084)	1.197	-0.383*** (0.088)	1.168
m40	-3.168*** (0.012)	-	-31.448*** (1.094)	-

Effect	Base model		NP Predictor	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m40*OTH	1.436*** (0.079)	2.898	23.105*** (4.159)	2.773
m41	0.293*** (0.013)	-	-0.018 (0.017)	-
m41*OTH	-0.032 (0.085)	1.400	0.201* (0.095)	1.368
m42	-0.502*** (0.013)	-	-2.457*** (0.076)	-
m42*OTH	0.111 (0.081)	1.546	1.626*** (0.296)	1.538
delta1	4.566*** (0.004)	-	4.648*** (0.004)	-
delta1*OTH	-0.954*** (0.024)	-	-0.956*** (0.025)	-
delta2	6.875*** (0.006)	-	7.136*** (0.006)	-
delta2*OTH	-1.265*** (0.050)	-	-1.337*** (0.052)	-
delta3	8.096*** (0.007)	-	8.449*** (0.008)	-
delta3*OTH	-1.260*** (0.085)	-	-1.403*** (0.086)	-
Intercept Variance	1.765		1.788	
NP Variance	-		0.113	
Intercept*Feature Covariance	-		0.394	

Note: + *p* < .10, * *p* < .05, ** *p* < .01, *** *p* < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G12.

EPvOTH Models' Adjusted DIF Estimates – Mathematics Assessment

	Base model		NP predictor	
Effect	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
m01*OTH	0.855* (0.079)	[0.700, 1.010]	0.851* (0.080)	[0.695, 1.008]
m02*OTH	1.371 (0.084)	[1.206, 1.536]	1.338 (0.084)	[1.174, 1.503]
m03*OTH	1.241* (0.082)	[1.080, 1.402]	1.227* (0.294)	[0.650, 1.803]
m04*OTH	1.472 (0.087)	[1.301, 1.643]	1.432 (0.085)	[1.265, 1.599]
m05*OTH	1.783* (0.090)	[1.607, 1.959]	1.797* (0.852)	[0.127, 3.467]
m06*OTH	0.916* (0.079)	[0.761, 1.071]	0.910* (0.080)	[0.754, 1.067]
m07*OTH	1.426 (0.099)	[1.232, 1.620]	1.387 (0.101)	[1.189, 1.585]
m08*OTH	1.263 (0.085)	[1.096, 1.430]	1.231 (0.088)	[1.059, 1.404]
m09*OTH	3.702* (0.081)	[3.543, 3.861]	3.879* (0.086)	[3.711, 4.048]
m10*OTH	1.056* (0.082)	[0.895, 1.217]	1.074* (1.221)	[-1.319, 3.468]
m11*OTH	1.038* (0.083)	[0.875, 1.201]	1.014* (0.086)	[0.846, 1.183]
m12*OTH	3.307* (0.082)	[3.146, 3.468]	3.372* (0.333)	[2.719, 4.025]
m13*OTH	-	-	-	-
m14*OTH	4.958* (0.080)	[4.801, 5.115]	5.050* (1.964)	[1.201, 8.900]
m15*OTH	1.112* (0.087)	[0.941, 1.283]	1.112* (0.938)	[-0.726, 2.951]
m16*OTH	2.587* (0.080)	[2.430, 2.744]	2.656* (0.646)	[1.390, 3.922]
m17*OTH	1.035* (0.083)	[0.872, 1.198]	1.012* (0.083)	[0.850, 1.175]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	NP predictor Adj. Estimate (SE)	NP predictor 95% CI
m18*OTH	2.021* (0.098)	[1.829, 2.213]	1.958* (0.096)	[1.770, 2.146]
m19*OTH	2.510* (0.081)	[2.351, 2.669]	2.555* (0.646)	[1.289, 3.821]
m20*OTH	0.599* (0.083)	[0.436, 0.762]	0.592* (0.093)	[0.410, 0.774]
m21*OTH	0.932* (0.082)	[0.771, 1.093]	0.910* (0.085)	[0.744, 1.077]
m22*OTH	0.937* (0.081)	[0.778, 1.096]	0.932* (0.452)	[0.046, 1.817]
m23*OTH	1.157* (0.079)	[1.002, 1.312]	1.175* (0.749)	[-0.293, 2.643]
m24*OTH	0.028* (0.081)	[-0.131, 0.187]	0.038* (0.085)	[-0.128, 0.205]
m25*OTH	1.514 (0.08)	[1.357, 1.671]	1.502 (0.085)	[1.335, 1.668]
m26*OTH	0.971* (0.081)	[0.812, 1.130]	0.951* (0.085)	[0.785, 1.118]
m27*OTH	1.337 (0.081)	[1.178, 1.496]	1.322 (0.194)	[0.942, 1.702]
m28*OTH	0.585* (0.082)	[0.424, 0.746]	0.579* (0.086)	[0.410, 0.747]
m29*OTH	2.125* (0.092)	[1.945, 2.305]	2.198* (1.355)	[-0.458, 4.854]
m30*OTH	5.150* (0.081)	[4.991, 5.309]	5.423* (0.144)	[5.140, 5.705]
m31*OTH	1.865* (0.081)	[1.706, 2.024]	1.900* (0.938)	[0.062, 3.739]
m32*OTH	1.058* (0.087)	[0.887, 1.229]	1.035* (0.090)	[0.858, 1.211]
m33*OTH	2.619* (0.081)	[2.460, 2.778]	2.661* (2.296)	[-1.839, 7.161]
m34*OTH	0.720* (0.081)	[0.561, 0.879]	0.709* (0.131)	[0.452, 0.965]
m35*OTH	5.753* (0.083)	[5.590, 5.916]	5.670* (5.075)	[-4.277, 15.617]

Effect	Base model		NP predictor	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
m36*OTH	1.386 (0.081)	[1.227, 1.545]	1.440 (1.593)	[-1.682, 4.562]
m37*OTH	2.463* (0.081)	[2.304, 2.622]	2.499* (0.514)	[1.492, 3.507]
m38*OTH	0.867* (0.080)	[0.710, 1.024]	0.918* (3.717)	[-6.368, 8.203]
m39*OTH	1.173* (0.084)	[1.008, 1.338]	1.144* (0.088)	[0.972, 1.317]
m40*OTH	2.840* (0.079)	[2.685, 2.995]	2.717* (4.159)	[-5.435, 10.868]
m41*OTH	1.372 (0.085)	[1.205, 1.539]	1.340 (0.095)	[1.154, 1.526]
m42*OTH	1.515 (0.081)	[1.356, 1.674]	1.507 (0.296)	[0.926, 2.087]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G13.

OTHvSPA Model Results – Mathematics Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
Intercept	0.719*** (0.062)	-	4.678*** (0.561)	-	1.918*** (0.496)	-	2.394*** (0.235)	-	3.296*** (0.834)	-
Intercept*SPA	0.463*** (0.081)	-	-2.551*** (0.641)	-	-0.638 (0.560)	-	-0.783** (0.269)	-	-0.848 (0.999)	-
LEX	-	-	3.723*** (0.515)	-	-	-	-	-	4.576*** (1.268)	-
LEX*SPA	-	-	-2.788*** (0.588)	-	-	-	-	-	-1.552 (1.541)	-
NP	-	-	-	-	1.938* (0.778)	-	-	-	-5.245*** (1.174)	-
NP*SPA	-	-	-	-	-1.749* (0.875)	-	-	-	2.223 (1.384)	-
RC	-	-	-	-	-	-	4.022*** (0.535)	-	2.385* (1.169)	-
RC*SPA	-	-	-	-	-	-	-2.94*** (0.606)	-	-2.450+ (1.408)	-
m01	-1.655*** (0.077)	-	-1.765*** (0.079)	-	-1.671*** (0.078)	-	-1.661*** (0.077)	-	-1.738*** (0.088)	-
m01*SPA	-0.100 (0.102)	0.370	0.026 (0.103)	0.041	-0.072 (0.102)	0.386	-0.079 (0.101)	0.588	-0.044 (0.113)	0.358
m02	0.169* (0.082)	-	-0.735*** (0.148)	-	0.136+ (0.082)	-	0.164* (0.081)	-	-0.872** (0.312)	-
m02*SPA	0.468*** (0.113)	0.950	1.133*** (0.180)	0.584	0.487*** (0.113)	0.956	0.462*** (0.112)	1.140	0.803* (0.383)	0.897

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m03	-0.229** (0.080)	-	-1.732*** (0.226)	-	-0.735*** (0.222)	-	-0.227** (0.079)	-	-0.690 (0.485)	-
m03*SPA	-0.045 (0.106)	0.427	1.104*** (0.263)	0.077	0.422+ (0.256)	0.438	-0.043 (0.104)	0.624	0.003 (0.586)	0.388
m04	0.336*** (0.084)	-	-4.574*** (0.699)	-	0.324*** (0.083)	-	0.327*** (0.083)	-	-5.758*** (1.710)	-
m04*SPA	0.164 (0.114)	0.640	3.916*** (0.800)	0.283	0.166 (0.113)	0.654	0.163 (0.113)	0.834	2.251 (2.078)	0.594
m05	0.705*** (0.088)	-	-1.037*** (0.255)	-	-0.803 (0.628)	-	0.688*** (0.086)	-	2.748** (0.872)	-
m05*SPA	0.202+ (0.121)	0.679	1.509*** (0.298)	0.331	1.599* (0.710)	0.688	0.204+ (0.119)	0.876	-0.845 (1.030)	0.643
m06	-1.487*** (0.077)	-	-1.589*** (0.079)	-	-1.500*** (0.078)	-	-1.487*** (0.077)	-	-1.561*** (0.087)	-
m06*SPA	0.016 (0.101)	0.489	0.131 (0.102)	0.148	0.041 (0.102)	0.501	0.030 (0.101)	0.699	0.061 (0.113)	0.466
m07	1.358*** (0.097)	-	0.987*** (0.100)	-	1.368*** (0.098)	-	1.326*** (0.095)	-	0.790*** (0.145)	-
m07*SPA	-0.191 (0.130)	0.278	0.065 (0.133)	0.025	-0.217+ (0.131)	0.313	-0.178 (0.128)	0.486	0.027 (0.185)	0.277
m08	0.205* (0.083)	-	-0.765*** (0.156)	-	0.252** (0.084)	-	0.200* (0.082)	-	-1.136** (0.356)	-
m08*SPA	-0.032 (0.110)	0.440	0.701*** (0.187)	0.092	-0.078 (0.111)	0.455	-0.030 (0.109)	0.638	0.440 (0.434)	0.402
m09	-1.504*** (0.079)	-	-11.97*** (1.45)	-	-1.418*** (0.082)	-	-1.487*** (0.079)	-	-14.528*** (3.574)	-
m09*SPA	0.992*** (0.105)	1.485	8.964*** (1.654)	1.284	0.928*** (0.107)	1.481	0.994*** (0.105)	1.683	5.535 (4.343)	1.571

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m10	-0.193* (0.080)	-	-8.410*** (1.160)	-	-2.381** (0.898)	-	-4.714*** (0.616)	-	-7.012*** (1.50)	-
m10*SPA	-0.141 (0.106)	0.329	6.109*** (1.324)	0.031	1.867+ (1.012)	0.337	3.208*** (0.700)	0.519	3.575* (1.795)	0.280
m11	-0.105 (0.081)	-	-1.128*** (0.164)	-	-0.049 (0.082)	-	-0.103 (0.080)	-	-1.518*** (0.379)	-
m11*SPA	-0.225* (0.106)	0.243	0.563** (0.194)	0.103	-0.271* (0.107)	0.258	-0.221* (0.105)	0.443	0.279 (0.461)	0.208
m12	-0.236** (0.080)	-	-7.934*** (1.083)	-	-0.809** (0.250)	-	-13.731*** (1.834)	-	-16.144*** (2.416)	-
m12*SPA	-0.245* (0.105)	0.222	5.593*** (1.236)	0.133	0.287 (0.287)	0.233	9.719*** (2.078)	0.319	10.647*** (2.870)	0.113
m13	-	-	-	-	-	-	-	-	-	-
m13*SPA	-	-	-	-	-	-	-	-	-	-
m14	-1.353*** (0.078)	-	-12.369*** (1.526)	-	-4.949*** (1.442)	-	-15.137*** (1.836)	-	-13.33*** (1.791)	-
m14*SPA	0.633*** (0.104)	1.119	9.000*** (1.741)	0.897	3.914* (1.625)	1.173	10.834*** (2.079)	1.457	9.657*** (2.048)	1.251
m15	0.486*** (0.085)	-	-2.191*** (0.385)	-	-1.178+ (0.691)	-	-3.973*** (0.617)	-	-0.849 (0.864)	-
m15*SPA	-0.367*** (0.111)	0.098	1.671*** (0.442)	0.249	1.167 (0.780)	0.101	2.944*** (0.701)	0.249	1.581 (1.001)	0.037
m16	-2.036*** (0.078)	-	-5.175*** (0.441)	-	-3.208*** (0.477)	-	-2.073*** (0.078)	-	-2.746** (0.997)	-
m16*SPA	0.475*** (0.102)	0.957	2.882*** (0.505)	0.668	1.548** (0.539)	0.984	0.533*** (0.102)	1.212	0.518 (1.201)	0.999
m17	-0.074 (0.081)	-	-0.208** (0.080)	-	-0.100 (0.08)	-	-0.074 (0.080)	-	-0.167+ (0.089)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
m17*SPA	0.183+ (0.108)	0.659	0.279** (0.107)	0.299	0.205+ (0.108)	0.668	0.181+ (0.107)	0.853	0.203+ (0.117)	0.611
m18	1.299*** (0.096)	-	-1.327*** (0.377)	-	1.257*** (0.095)	-	1.267*** (0.095)	-	-1.950* (0.905)	-
m18*SPA	0.130 (0.134)	0.605	2.113*** (0.437)	0.256	0.146 (0.133)	0.633	0.138 (0.133)	0.809	1.237 (1.101)	0.568
m19	-0.457*** (0.079)	-	-8.384*** (1.112)	-	-1.608*** (0.477)	-	-0.444*** (0.078)	-	-7.076** (2.506)	-
m19*SPA	-0.150 (0.104)	0.319	5.860*** (1.269)	0.020	0.908+ (0.540)	0.331	-0.150 (0.103)	0.515	1.863 (3.045)	0.297
m20	-0.048 (0.081)	-	-0.313*** (0.087)	-	-0.125 (0.086)	-	-0.045 (0.080)	-	-0.158 (0.116)	-
m20*SPA	-0.254* (0.106)	0.213	-0.038 (0.112)	0.130	-0.178 (0.111)	0.227	-0.251* (0.105)	0.412	-0.224 (0.146)	0.180
m21	-0.349*** (0.080)	-	-1.435*** (0.172)	-	-0.289*** (0.081)	-	-0.345*** (0.079)	-	-1.843*** (0.402)	-
m21*SPA	0.057 (0.105)	0.531	0.889*** (0.204)	0.176	0.006 (0.107)	0.540	0.058 (0.104)	0.727	0.583 (0.490)	0.488
m22	-0.422*** (0.079)	-	-2.056*** (0.244)	-	-1.217*** (0.336)	-	-0.415*** (0.078)	-	-0.245 (0.575)	-
m22*SPA	-0.445*** (0.103)	0.018	0.817** (0.283)	0.324	0.290 (0.382)	0.030	-0.437*** (0.102)	0.222	-0.670 (0.690)	0.012
m23	-1.167*** (0.078)	-	-7.945*** (0.948)	-	-2.511*** (0.552)	-	-1.161*** (0.077)	-	-5.838** (2.106)	-
m23*SPA	-0.039 (0.102)	0.433	5.093*** (1.082)	0.102	1.191+ (0.624)	0.445	-0.031 (0.101)	0.637	1.268 (2.556)	0.416
m24	-0.580*** (0.079)	-	-0.677*** (0.078)	-	-0.510*** (0.081)	-	-0.568*** (0.078)	-	-0.862*** (0.102)	-
m24*SPA	-0.396*** (0.103)	0.068	-0.287** (0.102)	0.279	-0.444*** (0.104)	0.081	-0.391*** (0.102)	0.269	-0.269* (0.129)	0.033

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
m25	-0.814*** (0.078)	-	-1.890*** (0.172)	-	-0.856*** (0.080)	-	-0.807*** (0.078)	-	-2.006*** (0.374)	-
m25*SPA	-0.311** (0.102)	0.155	0.531** (0.202)	0.189	-0.261* (0.104)	0.168	-0.304** (0.102)	0.358	0.101 (0.456)	0.123
m26	-0.501*** (0.079)	-	-1.228*** (0.131)	-	-0.435*** (0.081)	-	-0.492*** (0.078)	-	-1.556*** (0.286)	-
m26*SPA	-0.175+ (0.103)	0.294	0.398* (0.157)	0.058	-0.225* (0.105)	0.304	-0.172+ (0.102)	0.493	0.208 (0.350)	0.254
m27	-0.470*** (0.079)	-	-1.622*** (0.181)	-	-0.781*** (0.152)	-	-0.464*** (0.078)	-	-1.024** (0.372)	-
m27*SPA	0.108 (0.105)	0.583	0.990*** (0.213)	0.225	0.397* (0.179)	0.591	0.108 (0.103)	0.778	0.228 (0.453)	0.538
m28	-0.238** (0.080)	-	-7.646*** (1.045)	-	-0.283*** (0.082)	-	-0.231** (0.079)	-	-9.257*** (2.549)	-
m28*SPA	-0.110 (0.105)	0.360	5.524*** (1.192)	0.010	-0.062 (0.107)	0.371	-0.110 (0.104)	0.556	2.967 (3.098)	0.322
m29	0.871*** (0.089)	-	-2.617*** (0.500)	-	-1.510 (0.996)	-	0.851*** (0.088)	-	3.224* (1.466)	-
m29*SPA	-0.033 (0.121)	0.439	2.623*** (0.573)	0.082	2.182+ (1.123)	0.433	-0.024 (0.119)	0.644	-1.381 (1.739)	0.405
m30	-1.237*** (0.079)	-	-8.775*** (1.045)	-	-1.421*** (0.117)	-	-15.033*** (1.837)	-	-18.109*** (2.552)	-
m30*SPA	0.302** (0.104)	0.781	5.997*** (1.192)	0.492	0.490*** (0.142)	0.783	10.397*** (2.080)	1.011	11.635*** (3.037)	0.805
m31	-0.508*** (0.079)	-	-8.821*** (1.170)	-	-2.189** (0.690)	-	-0.501*** (0.078)	-	-6.165* (2.598)	-
m31*SPA	0.057 (0.104)	0.531	6.385*** (1.335)	0.197	1.600* (0.779)	0.543	0.057 (0.103)	0.726	1.629 (3.153)	0.509

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
m32	0.434*** (0.084)	-	-0.547*** (0.157)	-	0.370*** (0.086)	-	0.425*** (0.083)	-	-0.624+ (0.328)	-
m32*SPA	0.110 (0.114)	0.585	0.842*** (0.189)	0.236	0.162 (0.115)	0.599	0.110 (0.113)	0.780	0.456 (0.402)	0.546
m33	-2.932*** (0.079)	-	-9.625*** (0.930)	-	-7.089*** (1.686)	-	-3.036*** (0.079)	-	0.129 (2.580)	-
m33*SPA	0.412*** (0.104)	0.893	5.461*** (1.061)	0.583	4.168* (1.899)	0.872	0.516*** (0.104)	1.195	-1.522 (3.066)	0.946
m34	-0.355*** (0.079)	-	-2.334*** (0.290)	-	-0.533*** (0.109)	-	-0.348*** (0.079)	-	-2.290*** (0.651)	-
m34*SPA	-0.125 (0.104)	0.345	1.389*** (0.335)	0.005	0.046 (0.134)	0.356	-0.123 (0.103)	0.543	0.502 (0.792)	0.306
m35	-0.309*** (0.081)	-	-17.268*** (2.367)	-	-9.557* (3.726)	-	-14.054*** (1.836)	-	-4.182 (3.674)	-
m35*SPA	0.847*** (0.112)	1.337	13.973*** (2.700)	1.340	9.445* (4.195)	1.579	11.213*** (2.081)	1.844	6.130 (4.212)	1.731
m36	-0.426*** (0.079)	-	-4.616*** (0.594)	-	-3.290** (1.171)	-	-4.960*** (0.616)	-	-0.422 (1.184)	-
m36*SPA	-0.033 (0.104)	0.439	3.155*** (0.679)	0.093	2.593* (1.319)	0.449	3.325*** (0.700)	0.638	1.207 (1.356)	0.409
m37	-0.350*** (0.079)	-	-9.217*** (1.247)	-	-1.257*** (0.381)	-	-0.338*** (0.079)	-	-8.808** (2.864)	-
m37*SPA	-0.078 (0.105)	0.393	6.664*** (1.423)	0.055	0.758+ (0.432)	0.403	-0.080 (0.104)	0.586	2.624 (3.482)	0.373
m38	-0.946*** (0.078)	-	-6.395*** (0.766)	-	-7.682** (2.729)	-	-0.937*** (0.077)	-	10.739** (3.717)	-
m38*SPA	0.088 (0.102)	0.562	4.226*** (0.874)	0.230	6.235* (3.073)	0.586	0.091 (0.101)	0.761	-5.396 (4.370)	0.542
m39	0.155+ (0.082)	-	-2.246*** (0.347)	-	0.205* (0.084)	-	0.153+ (0.081)	-	-2.961*** (0.848)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
m39*SPA	-0.333** (0.108)	0.133	1.496*** (0.399)	0.215	-0.375*** (0.109)	0.151	-0.328** (0.107)	0.333	0.751 (1.031)	0.097
m40	-1.690*** (0.078)	-	-11.187*** (1.324)	-	-9.243** (3.053)	-	-6.283*** (0.616)	-	4.463 (3.591)	-
m40*SPA	0.198+ (0.102)	0.675	7.347*** (1.510)	0.325	7.025* (3.438)	0.648	3.569*** (0.699)	0.887	-1.737 (4.189)	0.650
m41	0.253** (0.083)	-	-0.993*** (0.191)	-	0.166+ (0.088)	-	0.247** (0.082)	-	-1.063* (0.415)	-
m41*SPA	-0.096 (0.110)	0.375	0.849*** (0.225)	0.030	-0.018 (0.115)	0.391	-0.093 (0.109)	0.573	0.340 (0.507)	0.340
m42	-0.380*** (0.079)	-	-6.496*** (0.862)	-	-0.887*** (0.223)	-	-0.374*** (0.078)	-	-6.530** (1.997)	-
m42*SPA	-0.298+ (0.104)	0.168	4.344*** (0.984)	0.185	0.172 (0.257)	0.180	-0.294** (0.103)	0.368	1.693 (2.428)	0.131
delta1	3.527*** (0.023)	-	3.698*** (0.025)	-	3.589*** (0.024)	-	3.657*** (0.025)	-	3.738*** (0.026)	-
delta1*SPA	-0.062+ (0.033)	-	-0.159*** (0.034)	-	-0.080* (0.033)	-	-0.135*** (0.034)	-	-0.171*** (0.035)	-
delta2	5.504*** (0.049)	-	5.879*** (0.052)	-	5.644*** (0.050)	-	5.856*** (0.053)	-	6.020*** (0.054)	-
delta2*SPA	0.366*** (0.084)	-	0.140 (0.087)	-	0.312*** (0.085)	-	0.174* (0.087)	-	0.085 (0.089)	-
delta3	6.727*** (0.085)	-	7.201*** (0.088)	-	6.880*** (0.085)	-	7.236*** (0.091)	-	7.440*** (0.093)	-
delta3*SPA	0.974*** (0.188)	-	0.672*** (0.191)	-	0.909*** (0.188)	-	0.695*** (0.195)	-	0.572** (0.197)	-
Intercept Variance	0.747		0.787		0.756		0.785		0.798	
LEX Variance	-		0.115		-		-		0.054	

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
NP Variance	-		-		0.03		-		0.028	
RC Variance	-		-		-		0.123		0.071	
Intercept*Feature Covariance	-		0.271		0.15		0.268		See Table G15	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G14.

OTHvSPA Models' Adjusted DIF Estimates – Mathematics Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
m01*SPA	0.363 (0.102)	[0.163, 0.563]	0.040* (0.103)	[-0.162, 0.242]	0.378* (0.102)	[0.178, 0.578]	0.576* (0.101)	[0.378, 0.774]	0.351* (0.113)	[0.130, 0.573]
m02*SPA	0.931* (0.113)	[0.710, 1.152]	0.573* (0.180)	[0.220, 0.925]	0.937* (0.113)	[0.715, 1.158]	1.117* (0.112)	[0.897, 1.336]	0.878* (0.383)	[0.128, 1.629]
m03*SPA	0.418 (0.106)	[0.210, 0.626]	0.075* (0.263)	[-0.440, 0.591]	0.429* (0.256)	[-0.072, 0.931]	0.612* (0.104)	[0.408, 0.816]	0.380* (0.586)	[-0.768, 1.529]
m04*SPA	0.627 (0.114)	[0.404, 0.850]	0.278* (0.80)	[-1.290, 1.846]	0.640* (0.113)	[0.419, 0.862]	0.818* (0.113)	[0.596, 1.039]	0.582 (2.078)	[-3.491, 4.655]
m05*SPA	0.665 (0.121)	[0.428, 0.902]	0.324* (0.298)	[-0.260, 0.908]	0.674 (0.710)	[-0.717, 2.066]	0.859* (0.119)	[0.625, 1.092]	0.630 (1.030)	[-1.389, 2.649]
m06*SPA	0.479 (0.101)	[0.281, 0.677]	0.145* (0.102)	[-0.055, 0.345]	0.491* (0.102)	[0.291, 0.691]	0.685* (0.101)	[0.487, 0.883]	0.456* (0.113)	[0.235, 0.678]
m07*SPA	0.272 (0.130)	[0.017, 0.527]	-0.024* (0.133)	[-0.285, 0.236]	0.306* (0.131)	[0.050, 0.563]	0.477* (0.128)	[0.226, 0.728]	0.271* (0.185)	[-0.091, 0.634]
m08*SPA	0.431 (0.110)	[0.215, 0.647]	0.090* (0.187)	[-0.276, 0.457]	0.445* (0.111)	[0.228, 0.663]	0.625* (0.109)	[0.411, 0.838]	0.394* (0.434)	[-0.456, 1.245]
m09*SPA	1.455* (0.105)	[1.249, 1.661]	1.258* (1.654)	[-1.984, 4.500]	1.451* (0.107)	[1.242, 1.661]	1.649* (0.105)	[1.443, 1.854]	1.539 (4.343)	[-6.973, 10.052]
m10*SPA	0.322 (0.106)	[0.114, 0.530]	-0.030 (1.324)	[-2.625, 2.565]	0.330 (1.012)	[-1.654, 2.314]	0.508 (0.700)	[-0.864, 1.880]	0.275 (1.795)	[-3.243, 3.793]
m11*SPA	0.238* (0.106)	[0.030, 0.446]	-0.101* (0.194)	[-0.481, 0.280]	0.252* (0.107)	[0.043, 0.462]	0.434* (0.105)	[0.228, 0.639]	0.204* (0.461)	[-0.700, 1.107]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
m12*SPA	0.218* (0.105)	[0.012, 0.424]	-0.131 (1.236)	[-2.553, 2.292]	0.228* (0.287)	[-0.335, 0.790]	0.313 (2.078)	[-3.760, 4.386]	0.111 (2.870)	[-5.514, 5.736]
m13*SPA	-	-	-	-	-	-	-	-	-	-
m14*SPA	1.096* (0.104)	[0.892, 1.300]	0.879* (1.741)	[-2.534, 4.291]	1.149 (1.625)	[-2.036, 4.334]	1.428 (2.079)	[-2.647, 5.503]	1.225 (2.048)	[-2.789, 5.240]
m15*SPA	0.096* (0.111)	[-0.122, 0.314]	-0.244* (0.442)	[-1.111, 0.622]	0.099 (0.780)	[-1.430, 1.628]	0.244 (0.701)	[-1.130, 1.618]	0.036 (1.001)	[-1.926, 1.998]
m16*SPA	0.938* (0.102)	[0.738, 1.138]	0.654* (0.505)	[-0.335, 1.644]	0.964* (0.539)	[-0.092, 2.021]	1.188* (0.102)	[0.988, 1.388]	0.979 (1.201)	[-1.375, 3.333]
m17*SPA	0.646 (0.108)	[0.434, 0.858]	0.293* (0.107)	[0.083, 0.503]	0.655* (0.108)	[0.443, 0.867]	0.836* (0.107)	[0.626, 1.045]	0.598* (0.117)	[0.369, 0.828]
m18*SPA	0.593 (0.134)	[0.330, 0.856]	0.251* (0.437)	[-0.606, 1.107]	0.620* (0.133)	[0.360, 0.881]	0.793* (0.133)	[0.532, 1.053]	0.557 (1.101)	[-1.601, 2.715]
m19*SPA	0.313 (0.104)	[0.109, 0.517]	-0.020* (1.269)	[-2.507, 2.467]	0.324 (0.540)	[-0.734, 1.383]	0.505* (0.103)	[0.303, 0.707]	0.291 (3.045)	[-5.677, 6.259]
m20*SPA	0.209* (0.106)	[0.001, 0.417]	-0.127* (0.112)	[-0.347, 0.092]	0.223* (0.111)	[0.005, 0.440]	0.404* (0.105)	[0.198, 0.609]	0.176* (0.146)	[-0.110, 0.462]
m21*SPA	0.520 (0.105)	[0.314, 0.726]	0.173* (0.204)	[-0.227, 0.572]	0.529* (0.107)	[0.320, 0.739]	0.713* (0.104)	[0.509, 0.917]	0.478* (0.490)	[-0.482, 1.439]
m22*SPA	0.018* (0.103)	[-0.184, 0.220]	-0.318* (0.283)	[-0.872, 0.237]	0.030 (0.382)	[-0.719, 0.779]	0.218* (0.102)	[0.018, 0.418]	-0.012 (0.690)	[-1.364, 1.341]
m23*SPA	0.424 (0.102)	[0.224, 0.624]	0.100* (1.082)	[-2.021, 2.220]	0.436 (0.624)	[-0.787, 1.659]	0.624* (0.101)	[0.426, 0.822]	0.407 (2.556)	[-4.602, 5.417]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
m24*SPA	0.067* (0.103)	[-0.135, 0.269]	-0.273* (0.102)	[-0.473, -0.073]	0.079* (0.104)	[-0.125, 0.283]	0.264* (0.102)	[0.064, 0.464]	0.033* (0.129)	[-0.220, 0.286]
m25*SPA	0.152* (0.102)	[-0.048, 0.352]	-0.185* (0.202)	[-0.581, 0.210]	0.164* (0.104)	[-0.039, 0.368]	0.351* (0.102)	[0.151, 0.551]	0.121* (0.456)	[-0.773, 1.014]
m26*SPA	0.288 (0.103)	[0.086, 0.490]	-0.056* (0.157)	[-0.364, 0.251]	0.298* (0.105)	[0.093, 0.504]	0.483* (0.102)	[0.283, 0.683]	0.249* (0.350)	[-0.437, 0.935]
m27*SPA	0.571 (0.105)	[0.365, 0.777]	0.221* (0.213)	[-0.197, 0.638]	0.579* (0.179)	[0.228, 0.930]	0.763* (0.103)	[0.561, 0.965]	0.527* (0.453)	[-0.361, 1.415]
m28*SPA	0.353 (0.105)	[0.147, 0.559]	0.009* (1.192)	[-2.327, 2.346]	0.363* (0.107)	[0.154, 0.573]	0.545* (0.104)	[0.341, 0.749]	0.316 (3.098)	[-5.756, 6.388]
m29*SPA	0.430 (0.121)	[0.193, 0.667]	0.080* (0.573)	[-1.043, 1.203]	0.425 (1.123)	[-1.776, 2.626]	0.631* (0.119)	[0.397, 0.864]	0.396 (1.739)	[-3.012, 3.805]
m30*SPA	0.765* (0.104)	[0.561, 0.969]	0.482* (1.192)	[-1.854, 2.819]	0.767* (0.142)	[0.488, 1.045]	0.991 (2.080)	[-3.086, 5.068]	0.789 (3.037)	[-5.164, 6.741]
m31*SPA	0.520 (0.104)	[0.316, 0.724]	0.193* (1.335)	[-2.424, 2.809]	0.532 (0.779)	[-0.995, 2.059]	0.712* (0.103)	[0.510, 0.914]	0.499 (3.153)	[-5.681, 6.679]
m32*SPA	0.573 (0.114)	[0.350, 0.796]	0.231* (0.189)	[-0.139, 0.602]	0.587* (0.115)	[0.362, 0.813]	0.765* (0.113)	[0.543, 0.986]	0.535* (0.402)	[-0.253, 1.323]
m33*SPA	0.875* (0.104)	[0.671, 1.079]	0.571* (1.061)	[-1.509, 2.650]	0.854 (1.899)	[-2.868, 4.576]	1.171* (0.104)	[0.967, 1.375]	0.927 (3.066)	[-5.082, 6.936]
m34*SPA	0.338 (0.104)	[0.134, 0.542]	-0.005* (0.335)	[-0.662, 0.652]	0.349* (0.134)	[0.086, 0.612]	0.532* (0.103)	[0.330, 0.734]	0.300 (0.792)	[-1.252, 1.852]
m35*SPA	1.310* (0.112)	[1.090, 1.530]	1.313 (2.700)	[-3.979, 6.605]	1.547 (4.195)	[-6.675, 9.769]	1.807 (2.081)	[-2.272, 5.886]	1.696 (4.212)	[-6.559, 9.952]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
m36*SPA	0.430 (0.104)	[0.226, 0.634]	0.091* (0.679)	[-1.240, 1.422]	0.440 (1.319)	[-2.145, 3.026]	0.625* (0.700)	[-0.747, 1.997]	0.401 (1.356)	[-2.257, 3.059]
m37*SPA	0.385 (0.105)	[0.179, 0.591]	0.054 (1.423)	[-2.735, 2.843]	0.395* (0.432)	[-0.452, 1.241]	0.575* (0.104)	[0.371, 0.779]	0.365 (3.482)	[-6.459, 7.190]
m38*SPA	0.551 (0.102)	[0.351, 0.751]	0.225* (0.874)	[-1.488, 1.938]	0.574 (3.073)	[-5.449, 6.597]	0.746* (0.101)	[0.548, 0.944]	0.531 (4.370)	[-8.034, 9.097]
m39*SPA	0.130* (0.108)	[-0.082, 0.342]	-0.210* (0.399)	[-0.992, 0.572]	0.148* (0.109)	[-0.065, 0.362]	0.327* (0.107)	[0.117, 0.536]	0.095 (1.031)	[-1.926, 2.116]
m40*SPA	0.661 (0.102)	[0.461, 0.861]	0.318 (1.510)	[-2.641, 3.278]	0.635 (3.438)	[-6.104, 7.373]	0.869* (0.699)	[-0.501, 2.239]	0.637 (4.189)	[-7.574, 8.847]
m41*SPA	0.367 (0.110)	[0.151, 0.583]	0.029* (0.225)	[-0.412, 0.470]	0.383* (0.115)	[0.158, 0.608]	0.562* (0.109)	[0.348, 0.775]	0.333* (0.507)	[-0.660, 1.327]
m42*SPA	0.165* (0.104)	[-0.039, 0.369]	-0.181* (0.984)	[-2.110, 1.748]	0.176* (0.257)	[-0.328, 0.680]	0.361* (0.103)	[0.159, 0.563]	0.128 (2.428)	[-4.630, 4.887]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G15.

Covariance Matrices for All Predictors Models – Mathematics Assessment

Comparison Group	Component	Intercept	LEX	NP	RC
STEBvLTEB	Intercept	0.846	**0.954**	**-0.066**	**0.696**
	LEX	0.212	0.058	**0.177**	**0.645**
	NP	-0.010	0.007	0.027	**-0.415**
	RC	0.167	0.041	-0.018	0.068
OTHvSPA	Intercept	0.798	**0.964**	**-0.105**	**0.694**
	LEX	0.201	0.054	**0.140**	**0.644**
	NP	-0.016	0.005	0.028	**-0.382**
	RC	0.165	0.040	-0.017	0.071

Note: Variances are on the diagonal, covariances are in the lower triangle, and correlations are in the upper triangle in bold.

Table G16.

EPvEB Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effec t Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	-0.997*** (0.020)	-	21.179*** (1.025)	-	15.833*** (1.212)	-	-8.505*** (0.430)	-	13.503*** (1.643)	-
Intercept*EB Status	1.307*** (0.057)	-	-14.949*** (1.519)	-	-8.701*** (2.049)	-	5.899*** (0.734)	-	-10.906*** (3.031)	-
LEX	-	-	13.684*** (0.633)	-	-	-	-	-	8.556*** (0.788)	-
LEX*EB Status	-	-	-9.997*** (0.941)	-	-	-	-	-	-6.939*** (1.305)	-
NP	-	-	-	-	20.613*** (1.485)	-	-	-	5.241*** (1.505)	-
NP*EB Status	-	-	-	-	-12.223*** (2.515)	-	-	-	-3.741 (2.678)	-
RC	-	-	-	-	-	-	-28.801*** (1.647)	-	-14.071*** (1.745)	-
RC*SPA	-	-	-	-	-	-	17.746*** (2.787)	-	8.313* (3.410)	-
b01	-0.369*** (0.027)	-	-53.043*** (2.432)	-	-23.140*** (1.640)	-	72.597*** (4.170)	-	-3.449 (6.390)	-
b01*EB Status	0.263*** (0.073)	1.602	38.805*** (3.616)	2.870	13.776*** (2.778)	0.626	-44.719*** (7.058)	0.805	10.076 (12.351)	1.789
b02	-	-	-	-	-	-	-	-	-	-
b02*EB Status	-	-	-	-	-	-	-	-	-	-
b03	0.344*** (0.026)	-	-24.513*** (1.149)	-	-21.557*** (1.578)	-	-1.960*** (0.134)	-	-21.925*** (1.605)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effec t Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b03*EB Status	-0.320*** (0.073)	1.007	17.866*** (1.708)	2.181	12.676*** (2.672)	0.028	1.100*** (0.235)	0.261	16.984*** (2.807)	1.194
b04	0.079** (0.026)	-	-27.716*** (1.284)	-	-8.439*** (0.614)	-	-2.226*** (0.134)	-	-20.63*** (1.395)	-
b04*EB Status	-0.078 (0.073)	1.254	20.258*** (1.910)	2.439	4.976*** (1.041)	0.265	1.342*** (0.235)	0.508	16.294*** (2.291)	1.452
b05	0.478*** (0.026)	-	-24.378*** (1.149)	-	-4.784*** (0.380)	-	-1.825*** (0.134)	-	-17.554*** (1.256)	-
b05*EB Status	-0.314*** (0.074)	1.013	17.871*** (1.708)	2.186	2.809*** (0.645)	0.024	1.105*** (0.235)	0.266	13.963*** (2.043)	1.192
b06	-0.076** (0.027)	-	-10.318*** (0.474)	-	-23.176*** (1.664)	-	-2.382*** (0.134)	-	-13.502*** (1.520)	-
b06*EB Status	-0.323*** (0.073)	1.004	7.177*** (0.707)	2.148	13.383*** (2.818)	0.026	1.097*** (0.235)	0.258	9.768*** (2.716)	1.157
b07	0.060* (0.026)	-	-11.637*** (0.541)	-	-6.395*** (0.466)	-	-2.245*** (0.134)	-	-10.040*** (0.587)	-
b07*EB Status	-0.564*** (0.073)	0.758	8.002*** (0.806)	1.909	3.269*** (0.791)	0.230	0.855*** (0.235)	0.011	7.241*** (0.966)	0.908
b08	-0.952*** (0.028)	-	-30.214*** (1.351)	-	-3.592*** (0.192)	-	-2.998*** (0.120)	-	-20.969*** (1.556)	-
b08*EB Status	0.232** (0.073)	1.571	21.637*** (2.009)	2.765	1.802*** (0.330)	0.581	1.492*** (0.211)	0.824	16.206*** (2.538)	1.776
b09	-0.005 (0.026)	-	-43.886*** (2.027)	-	-13.452*** (0.969)	-	22.758*** (1.301)	-	-19.777*** (3.266)	-
b09*EB Status	0.169* (0.074)	1.506	32.284*** (3.013)	2.745	8.149*** (1.641)	0.522	-13.865*** (2.203)	0.745	18.372** (6.021)	1.734
b10	0.909*** (0.026)	-	-42.909*** (2.027)	-	-2.955*** (0.279)	-	23.924*** (1.316)	-	-16.245*** (3.341)	-
b10*EB Status	-0.178* (0.077)	1.152	31.903*** (3.014)	2.356	2.113*** (0.476)	0.162	-14.367*** (2.229)	0.395	16.176** (6.070)	1.345

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effec t Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b11	-0.296*** (0.027)	-	-4.677*** (0.204)	-	-8.689*** (0.605)	-	-2.602*** (0.134)	-	-6.304*** (0.541)	-
b11*EB Status	0.445*** (0.074)	1.788	3.644*** (0.310)	2.899	5.425*** (1.026)	0.798	1.867*** (0.235)	1.044	4.863*** (0.953)	1.897
b12	-3.937*** (0.026)	-	-47.720*** (2.027)	-	-32.774*** (2.078)	-	-6.231*** (0.134)	-	-39.748*** (2.491)	-
b12*EB Status	4.179*** (0.075)	5.599	36.177*** (3.014)	6.718	21.284*** (3.519)	4.609	5.590*** (0.235)	4.843	32.278*** (4.298)	5.693
b13	-0.303*** (0.027)	-	-13.466*** (0.609)	-	-2.941*** (0.192)	-	-2.609*** (0.134)	-	-10.350*** (0.641)	-
b13*EB Status	0.270*** (0.073)	1.609	9.900*** (0.907)	2.754	1.835*** (0.330)	0.615	1.692*** (0.235)	0.865	8.121*** (1.019)	1.755
b14	-0.267*** (0.027)	-	-26.601*** (1.216)	-	-15.694*** (1.111)	-	23.737*** (1.372)	-	-8.942*** (2.562)	-
b14*EB Status	-0.121+ (0.073)	1.210	19.144*** (1.809)	2.394	9.034*** (1.883)	0.228	-14.918*** (2.323)	0.449	9.137+ (4.905)	1.372
b15	0.012 (0.026)	-	-16.073*** (0.743)	-	-0.606*** (0.052)	-	-2.293*** (0.134)	-	-11.351*** (0.850)	-
b15*EB Status	0.003 (0.073)	1.337	11.772*** (1.107)	2.492	0.370*** (0.105)	0.342	1.424*** (0.235)	0.591	8.975*** (1.363)	1.492
b16	0.163*** (0.026)	-	-30.551*** (1.419)	-	-22.152*** (1.607)	-	-2.142*** (0.134)	-	-25.883*** (1.798)	-
b16*EB Status	0.022 (0.074)	1.356	22.497*** (2.110)	2.551	13.264*** (2.722)	0.378	1.443*** (0.235)	0.611	20.396*** (3.110)	1.575
b17	1.336*** (0.026)	-	-40.987*** (1.959)	-	-67.202*** (4.940)	-	71.908*** (4.037)	-	-8.113 (7.199)	-
b17*EB Status	-0.659*** (0.076)	0.661	30.315*** (2.913)	1.827	40.007*** (8.366)	0.321	-44.161*** (6.833)	0.092	12.937 (13.939)	0.828
b18	0.114*** (0.026)	-	-7.198*** (0.339)	-	-17.252*** (1.251)	-	-2.191*** (0.134)	-	-10.017*** (1.141)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effec t Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
b18*EB Status	-0.084 (0.073)	1.248	5.265*** (0.507)	2.380	10.220*** (2.119)	0.265	1.336*** (0.235)	0.502	7.467*** (2.033)	1.385
b19	0.350*** (0.026)	-	0.333*** (0.025)	-	-0.271*** (0.052)	-	-1.953*** (0.134)	-	-0.954*** (0.146)	-
b19*EB Status	-0.445*** (0.073)	0.880	-0.425*** (0.071)	2.011	-0.074 (0.105)	0.111	0.974*** (0.235)	0.132	0.358 (0.284)	1.005
b20	0.517*** (0.026)	-	-8.274*** (0.406)	-	-2.745*** (0.236)	-	-1.785*** (0.134)	-	-6.953*** (0.406)	-
b20*EB Status	-0.457*** (0.073)	0.868	5.977*** (0.607)	2.015	1.480*** (0.404)	0.122	0.961*** (0.235)	0.119	5.286*** (0.640)	1.013
b21	0.679*** (0.026)	-	-13.962*** (0.676)	-	-70.198*** (5.106)	-	-1.622*** (0.134)	-	-27.646*** (4.881)	-
b21*EB Status	-0.654*** (0.073)	0.666	10.061*** (1.007)	1.827	41.394*** (8.647)	0.303	0.763** (0.235)	0.083	20.348* (8.736)	0.838
b22	0.887*** (0.026)	-	-7.917*** (0.406)	-	0.258*** (0.051)	-	-1.413*** (0.134)	-	-5.926*** (0.438)	-
b22*EB Status	-0.529*** (0.074)	0.794	5.912*** (0.607)	1.949	-0.155 (0.106)	0.194	0.888*** (0.235)	0.044	4.742*** (0.686)	0.947
b23	-3.316*** (0.027)	-	-35.461*** (1.486)	-	-20.787*** (1.258)	-	-4.516*** (0.074)	-	-28.467*** (1.740)	-
b23*EB Status	3.251*** (0.075)	4.652	26.757*** (2.210)	5.817	13.621*** (2.131)	3.674	3.989*** (0.139)	3.898	23.109*** (2.973)	4.813
b24	-0.228*** (0.027)	-	-30.951*** (1.419)	-	-35.786*** (2.561)	-	-1.439*** (0.074)	-	-29.126*** (2.459)	-
b24*EB Status	0.076 (0.073)	1.411	22.553*** (2.110)	2.608	21.176*** (4.337)	0.444	0.822*** (0.138)	0.665	22.544*** (4.391)	1.639
b25	-0.475*** (0.027)	-	-13.637*** (0.609)	-	-8.228*** (0.559)	-	-2.781*** (0.135)	-	-11.825*** (0.680)	-
b25*EB Status	0.626*** (0.074)	1.973	10.253*** (0.907)	3.115	5.226*** (0.949)	0.982	2.049*** (0.235)	1.229	9.407*** (1.131)	2.121

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effec t Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b26	0.463*** (0.026)	-	-5.409*** (0.272)	-	-0.160** (0.051)	-	-1.839*** (0.134)	-	-4.509*** (0.280)	-
b26*EB Status	-0.237*** (0.074)	1.092	4.057*** (0.408)	2.229	0.133 (0.105)	0.100	1.182*** (0.235)	0.344	3.536*** (0.431)	1.224
b27	-0.377*** (0.027)	-	-42.806*** (1.959)	-	-8.254*** (0.568)	-	-2.136*** (0.104)	-	-29.834*** (2.217)	-
b27*EB Status	-0.108 (0.073)	1.224	30.924*** (2.913)	2.448	4.568*** (0.964)	0.235	0.975*** (0.185)	0.477	23.433*** (3.652)	1.468
b28	-0.672*** (0.027)	-	-22.619*** (1.014)	-	-46.443*** (3.296)	-	95.177*** (5.478)	-	20.863** (7.224)	-
b28*EB Status	0.889*** (0.074)	2.241	16.949*** (1.508)	3.419	28.055*** (5.582)	1.290	-58.211*** (9.271)	1.416	-7.427 (14.193)	2.268
b29	-0.491*** (0.027)	-	-6.339*** (0.272)	-	-8.864*** (0.604)	-	-2.798*** (0.135)	-	-7.413*** (0.537)	-
b29*EB Status	0.016 (0.073)	1.350	4.303*** (0.409)	2.480	4.986*** (1.024)	0.363	1.437*** (0.235)	0.605	5.193*** (0.940)	1.480
b30	0.198*** (0.026)	-	-23.208*** (1.082)	-	-8.176*** (0.604)	-	-2.106*** (0.134)	-	-17.724*** (1.166)	-
b30*EB Status	-0.554*** (0.073)	0.769	16.569*** (1.609)	1.939	4.417*** (1.024)	0.218	0.865*** (0.235)	0.021	13.550*** (1.917)	0.944
b31	-0.105*** (0.027)	-	-19.126*** (0.879)	-	-2.744*** (0.192)	-	-2.411*** (0.134)	-	-13.823*** (0.968)	-
b31*EB Status	-0.088 (0.073)	1.244	13.828*** (1.308)	2.407	1.479*** (0.33)	0.251	1.332*** (0.235)	0.498	10.747*** (1.559)	1.411
b32	-3.303*** (0.027)	-	-31.100*** (1.284)	-	-24.481*** (1.526)	-	-5.596*** (0.135)	-	-27.216*** (1.659)	-
b32*EB Status	3.294*** (0.076)	4.696	23.625*** (1.910)	5.875	15.860*** (2.584)	3.714	4.702*** (0.235)	3.937	21.934*** (2.876)	4.864
b33	1.248*** (0.026)	-	-7.571*** (0.406)	-	0.615*** (0.051)	-	-1.049*** (0.134)	-	-5.581*** (0.438)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effec t Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b33*EB Status	-0.595*** (0.076)	0.727	5.853*** (0.607)	1.889	-0.218* (0.107)	0.258	0.821*** (0.236)	0.024	4.684*** (0.686)	0.888
b34	-0.405*** (0.027)	-	-9.176*** (0.406)	-	-3.043*** (0.192)	-	-2.164*** (0.104)	-	-7.430*** (0.416)	-
b34*EB Status	0.206** (0.073)	1.544	6.626*** (0.607)	2.678	1.773*** (0.330)	0.551	1.290*** (0.185)	0.799	5.664*** (0.662)	1.675
b35	0.591*** (0.026)	-	-31.561*** (1.486)	-	-6.507*** (0.512)	-	-1.711*** (0.134)	-	-22.468*** (1.641)	-
b35*EB Status	-0.391*** (0.074)	0.935	23.132*** (2.210)	2.118	3.821*** (0.868)	0.053	1.027*** (0.235)	0.186	17.930*** (2.685)	1.126
b36	-0.234*** (0.027)	-	-29.489*** (1.351)	-	-37.876*** (2.711)	-	72.989*** (4.185)	-	7.707 (5.895)	-
b36*EB Status	0.504*** (0.074)	1.848	21.915*** (2.009)	3.049	22.844*** (4.591)	0.887	-44.646*** (7.083)	1.043	1.001 (11.566)	1.939
b37	-0.335*** (0.027)	-	-17.888*** (0.811)	-	-9.450*** (0.657)	-	-1.546*** (0.074)	-	-14.248*** (0.925)	-
b37*EB Status	0.135+ (0.073)	1.472	12.979*** (1.207)	2.632	5.545*** (1.114)	0.484	0.881*** (0.138)	0.726	11.086*** (1.564)	1.638
b38	0.050+ (0.026)	-	-18.970*** (0.879)	-	-14.098*** (1.019)	-	0.050+ (0.026)	-	-15.468*** (1.166)	-
b38*EB Status	-0.010 (0.073)	1.324	13.906*** (1.308)	2.486	8.386*** (1.727)	0.340	-0.010 (0.074)	0.577	12.251*** (2.049)	1.494
b39	-0.063* (0.027)	-	-7.370*** (0.339)	-	-0.062* (0.026)	-	-2.369*** (0.134)	-	-5.768*** (0.376)	-
b39*EB Status	-0.514*** (0.073)	0.809	4.843*** (0.507)	1.950	-0.511*** (0.072)	0.183	0.905*** (0.235)	0.062	3.886*** (0.590)	0.945
b40	-0.308*** (0.027)	-	-17.861*** (0.811)	-	-46.841*** (3.351)	-	-0.425*** (0.028)	-	-23.213*** (3.130)	-
b40*EB Status	0.212** (0.073)	1.550	13.055*** (1.207)	2.709	27.825*** (5.675)	0.594	0.284*** (0.074)	0.804	17.656** (5.639)	1.740

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effec t Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b41	0.279*** (0.026)	-	-14.356*** (0.676)	-	-4.879*** (0.372)	-	-1.477*** (0.104)	-	-11.063*** (0.720)	-
b41*EB Status	-0.380*** (0.073)	0.946	10.329*** (1.007)	2.100	2.682*** (0.633)	0.043	0.701*** (0.185)	0.198	8.522*** (1.177)	1.102
b42	-0.536*** (0.027)	-	-38.577*** (1.757)	-	-72.833*** (5.206)	-	-2.842*** (0.135)	-	-43.909*** (4.799)	-
b42*EB Status	0.542*** (0.074)	1.887	28.380*** (2.612)	3.117	43.448*** (8.816)	0.958	1.964*** (0.235)	1.143	33.758*** (8.630)	2.187
b43	0.355*** (0.026)	-	-23.050*** (1.082)	-	-15.669*** (1.154)	-	23.663*** (1.333)	-	-6.982** (2.427)	-
b43*EB Status	-0.533*** (0.073)	0.790	16.589*** (1.609)	1.959	8.976*** (1.956)	0.193	-14.900*** (2.256)	0.032	7.549 (4.668)	0.946
b44	-3.435*** (0.027)	-	-50.116*** (2.162)	-	-26.691*** (1.676)	-	-2.879*** (0.041)	-	-38.246*** (2.561)	-
b44*EB Status	3.363*** (0.075)	4.766	37.468*** (3.215)	5.862	17.161*** (2.838)	3.782	3.019*** (0.092)	4.012	31.085*** (4.406)	4.842
b45	-5.112*** (0.026)	-	-32.934*** (1.284)	-	-25.263*** (1.451)	-	-5.770*** (0.046)	-	-28.014*** (1.665)	-
b45*EB Status	3.401*** (0.077)	4.805	23.746*** (1.910)	5.999	15.359*** (2.458)	3.826	3.801*** (0.100)	4.050	21.407*** (2.908)	5.000
delta1	5.358*** (0.010)	-	5.386*** (0.010)	-	5.362*** (0.010)	-	5.358*** (0.010)	-	5.405*** (0.010)	-
delta1*EB Status	-1.138*** (0.030)	-	-1.149*** (0.030)	-	-1.143*** (0.030)	-	-1.133*** (0.030)	-	-1.145*** (0.031)	-
delta2	6.977*** (0.014)	-	7.049*** (0.014)	-	6.985*** (0.014)	-	6.984*** (0.014)	-	7.104*** (0.014)	-
delta2*EB Status	-0.942*** (0.068)	-	-0.987*** (0.068)	-	-0.950*** (0.068)	-	-0.942*** (0.068)	-	-1.008*** (0.068)	-
delta3	8.320*** (0.019)	-	8.415*** (0.019)	-	8.331*** (0.019)	-	8.332*** (0.019)	-	8.492*** (0.019)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
delta3*EB Status	-1.103*** (0.120)	-	-1.166*** (0.120)	-	-1.112*** (0.120)	-	-1.106*** (0.120)	-	-1.206*** (0.121)	-
Intercept Variance	1.025		1.015		1.014		1.005		1.019	
LEX Variance	-		0.029		-		-		0.051	
NP Variance	-		-		0.006		-		0.005	
RC Variance	-		-		-		0.006		0.039	
Intercept*Feature Covariance	-		0.169		0.067		-0.076		See Table G30	

Note: + p < .10, * p < .05, ** p < .01, *** p < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G17.

EPvEB Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b01*EB Status	1.570* (0.073)	[1.427, 1.713]	2.812* (3.616)	[-4.275, 9.900]	0.614* (2.778)	[-4.831, 6.058]	0.789 (7.058)	[-13.045, 14.623]	1.753 (12.351)	[-22.455, 25.961]
b02*EB Status	-	-	-	-	-	-	-	-	-	-
b03*EB Status	0.987* (0.073)	[0.844, 1.130]	2.137* (1.708)	[-1.210, 5.485]	0.027* (2.672)	[-5.210, 5.264]	0.256* (0.235)	[-0.205, 0.716]	1.169* (2.807)	[-4.332, 6.671]
b04*EB Status	1.229 (0.073)	[1.086, 1.372]	2.390* (1.910)	[-1.354, 6.133]	0.260* (1.041)	[-1.781, 2.300]	0.498* (0.235)	[0.037, 0.958]	1.422* (2.291)	[-3.068, 5.913]
b05*EB Status	0.993* (0.074)	[0.848, 1.138]	2.142* (1.708)	[-1.205, 5.490]	0.024* (0.645)	[-1.240, 1.288]	0.261* (0.235)	[-0.200, 0.721]	1.167* (2.043)	[-2.837, 5.172]
b06*EB Status	0.984* (0.073)	[0.841, 1.127]	2.105* (0.707)	[0.719, 3.491]	0.025* (2.818)	[-5.498, 5.548]	0.253* (0.235)	[-0.208, 0.713]	1.133* (2.716)	[-4.190, 6.457]
b07*EB Status	0.743* (0.073)	[0.600, 0.886]	1.870* (0.806)	[0.291, 3.450]	-0.225* (0.791)	[-1.775, 1.325]	0.011* (0.235)	[-0.450, 0.471]	0.890* (0.966)	[-1.003, 2.783]
b08*EB Status	1.539* (0.073)	[1.396, 1.682]	2.709* (2.009)	[-1.228, 6.647]	0.569* (0.330)	[-0.078, 1.216]	0.807* (0.211)	[0.394, 1.221]	1.740* (2.538)	[-3.235, 6.714]
b09*EB Status	1.476* (0.074)	[1.331, 1.621]	2.689* (3.013)	[-3.216, 8.595]	0.511* (1.641)	[-2.705, 3.728]	0.730* (2.203)	[-3.588, 5.047]	1.699* (6.021)	[-10.102, 13.500]
b10*EB Status	1.129* (0.077)	[0.978, 1.280]	2.308* (3.014)	[-3.599, 8.216]	0.159* (0.476)	[-0.774, 1.092]	0.387* (2.229)	[-3.982, 4.756]	1.318* (6.070)	[-10.580, 13.215]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
b11*EB Status	1.752* (0.074)	[1.607, 1.897]	2.841* (0.310)	[2.233, 3.448]	0.782* (1.026)	[-1.229, 2.793]	1.023* (0.235)	[0.562, 1.483]	1.859* (0.953)	[-0.009, 3.727]
b12*EB Status	5.486* (0.075)	[5.339, 5.633]	6.582* (3.014)	[0.675, 12.490]	4.516* (3.519)	[-2.381, 11.413]	4.746* (0.235)	[4.285, 5.206]	5.578* (4.298)	[-2.846, 14.002]
b13*EB Status	1.577* (0.073)	[1.434, 1.720]	2.699* (0.907)	[0.921, 4.476]	0.602* (0.330)	[-0.045, 1.249]	0.848* (0.235)	[0.387, 1.308]	1.720* (1.019)	[-0.278, 3.717]
b14*EB Status	1.186 (0.073)	[1.043, 1.329]	2.346* (1.809)	[-1.200, 5.891]	0.223* (1.883)	[-3.468, 3.914]	0.440* (2.323)	[-4.113, 4.993]	1.344* (4.905)	[-8.269, 10.958]
b15*EB Status	1.310 (0.073)	[1.167, 1.453]	2.441* (1.107)	[0.272, 4.611]	0.335* (0.105)	[0.129, 0.541]	0.580* (0.235)	[0.119, 1.040]	1.462* (1.363)	[-1.209, 4.134]
b16*EB Status	1.329 (0.074)	[1.184, 1.474]	2.500* (2.110)	[-1.636, 6.635]	0.371* (2.722)	[-4.965, 5.706]	0.599* (0.235)	[0.138, 1.059]	1.544* (3.110)	[-4.552, 7.639]
b17*EB Status	0.648* (0.076)	[0.499, 0.797]	1.790* (2.913)	[-3.919, 7.500]	-0.315 (8.366)	[-16.712, 16.082]	-0.090 (6.833)	[-13.483, 13.302]	0.811 (13.939)	[-26.509, 28.132]
b18*EB Status	1.223 (0.073)	[1.080, 1.366]	2.332* (0.507)	[1.339, 3.326]	0.260* (2.119)	[-3.893, 4.413]	0.492* (0.235)	[0.031, 0.952]	1.357* (2.033)	[-2.627, 5.342]
b19*EB Status	0.862* (0.073)	[0.719, 1.005]	1.971* (0.071)	[1.832, 2.110]	-0.109* (0.105)	[-0.315, 0.097]	0.130* (0.235)	[-0.331, 0.590]	0.985* (0.284)	[0.428, 1.541]
b20*EB Status	0.850* (0.073)	[0.707, 0.993]	1.975* (0.607)	[0.785, 3.164]	-0.119* (0.404)	[-0.911, 0.672]	0.117* (0.235)	[-0.344, 0.577]	0.993* (0.640)	[-0.262, 2.247]
b21*EB Status	0.653* (0.073)	[0.510, 0.796]	1.790* (1.007)	[-0.184, 3.764]	-0.297 (8.647)	[-17.245, 16.651]	-0.081* (0.235)	[-0.542, 0.379]	0.821 (8.736)	[-16.301, 17.944]
b22*EB Status	0.778* (0.074)	[0.633, 0.923]	1.910* (0.607)	[0.720, 3.099]	-0.190* (0.106)	[-0.398, 0.018]	0.044* (0.235)	[-0.417, 0.504]	0.928* (0.686)	[-0.417, 2.272]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b23*EB Status	4.558* (0.075)	[4.411, 4.705]	5.700* (2.210)	[1.368, 10.031]	3.600* (2.131)	[-0.577, 7.777]	3.819* (0.139)	[3.546, 4.091]	4.716* (2.973)	[-1.111, 10.543]
b24*EB Status	1.383 (0.073)	[1.240, 1.526]	2.556* (2.110)	[-1.580, 6.691]	0.435* (4.337)	[-8.065, 8.936]	0.652* (0.138)	[0.381, 0.922]	1.606* (4.391)	[-7.000, 10.212]
b25*EB Status	1.933* (0.074)	[1.788, 2.078]	3.052* (0.907)	[1.274, 4.829]	0.962* (0.949)	[-0.898, 2.822]	1.205* (0.235)	[0.744, 1.665]	2.078* (1.131)	[-0.139, 4.295]
b26*EB Status	1.070* (0.074)	[0.925, 1.215]	2.184* (0.408)	[1.384, 2.984]	0.098* (0.105)	[-0.108, 0.304]	0.338* (0.235)	[-0.123, 0.798]	1.200* (0.431)	[0.355, 2.044]
b27*EB Status	1.199 (0.073)	[1.056, 1.342]	2.399* (2.913)	[-3.310, 8.109]	0.231* (0.964)	[-1.659, 2.120]	0.468* (0.185)	[0.105, 0.830]	1.438* (3.652)	[-5.720, 8.596]
b28*EB Status	2.196* (0.074)	[2.051, 2.341]	3.350* (1.508)	[0.394, 6.305]	1.264 (5.582)	[-9.677, 12.205]	1.387 (9.271)	[-16.784, 19.559]	2.222 (14.193)	[-25.596, 30.041]
b29*EB Status	1.323 (0.073)	[1.180, 1.466]	2.430* (0.409)	[1.628, 3.232]	0.355* (1.024)	[-1.652, 2.362]	0.593* (0.235)	[0.132, 1.053]	1.450* (0.940)	[-0.392, 3.292]
b30*EB Status	0.753* (0.073)	[0.610, 0.896]	1.900* (1.609)	[-1.254, 5.054]	-0.214* (1.024)	[-2.221, 1.793]	0.021* (0.235)	[-0.440, 0.481]	0.925* (1.917)	[-2.832, 4.682]
b31*EB Status	1.219 (0.073)	[1.076, 1.362]	2.358* (1.308)	[-0.206, 4.922]	0.246* (0.330)	[-0.401, 0.893]	0.488* (0.235)	[0.027, 0.948]	1.383* (1.559)	[-1.673, 4.438]
b32*EB Status	4.601* (0.076)	[4.452, 4.750]	5.757* (1.910)	[2.013, 9.500]	3.639* (2.584)	[-1.426, 8.703]	3.858* (0.235)	[3.397, 4.318]	4.765* (2.876)	[-0.871, 10.402]
b33*EB Status	0.712* (0.076)	[0.563, 0.861]	1.851* (0.607)	[0.661, 3.040]	-0.253* (0.107)	[-0.463, -0.043]	-0.023* (0.236)	[-0.486, 0.439]	0.870* (0.686)	[-0.475, 2.214]
b34*EB Status	1.513* (0.073)	[1.370, 1.656]	2.624* (0.607)	[1.434, 3.813]	0.540* (0.330)	[-0.107, 1.187]	0.783* (0.185)	[0.420, 1.145]	1.641* (0.662)	[0.343, 2.938]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b35*EB Status	0.916* (0.074)	[0.771, 1.061]	2.075* (2.210)	[-2.257, 6.406]	-0.052* (0.868)	[-1.753, 1.649]	0.183* (0.235)	[-0.278, 0.643]	1.103* (2.685)	[-4.160, 6.366]
b36*EB Status	1.811* (0.074)	[1.666, 1.956]	2.987* (2.009)	[-0.950, 6.925]	0.869* (4.591)	[-8.130, 9.867]	1.022 (7.083)	[-12.861, 14.904]	1.900 (11.566)	[-20.769, 24.569]
b37*EB Status	1.442 (0.073)	[1.299, 1.585]	2.579* (1.207)	[0.213, 4.944]	0.474* (1.114)	[-1.709, 2.658]	0.711* (0.138)	[0.440, 0.981]	1.605* (1.564)	[-1.460, 4.671]
b38*EB Status	1.297 (0.073)	[1.154, 1.440]	2.436* (1.308)	[-0.128, 5.000]	0.333* (1.727)	[-3.052, 3.718]	0.565* (0.074)	[0.420, 0.710]	1.464* (2.049)	[-2.552, 5.480]
b39*EB Status	0.793* (0.073)	[0.650, 0.936]	1.910* (0.507)	[0.917, 2.904]	-0.179* (0.072)	[-0.320, -0.038]	0.061* (0.235)	[-0.400, 0.521]	0.926* (0.590)	[-0.230, 2.083]
b40*EB Status	1.519* (0.073)	[1.376, 1.662]	2.655* (1.207)	[0.289, 5.020]	0.582 (5.675)	[-10.541, 11.705]	0.788* (0.074)	[0.643, 0.933]	1.705* (5.639)	[-9.347, 12.757]
b41*EB Status	0.927* (0.073)	[0.784, 1.070]	2.058* (1.007)	[0.084, 4.032]	-0.042* (0.633)	[-1.283, 1.199]	0.194* (0.185)	[-0.169, 0.556]	1.080* (1.177)	[-1.227, 3.387]
b42*EB Status	1.849* (0.074)	[1.704, 1.994]	3.054* (2.612)	[-2.065, 8.174]	0.938 (8.816)	[-16.341, 18.218]	1.120* (0.235)	[0.659, 1.580]	2.143 (8.630)	[-14.772, 19.058]
b43*EB Status	0.774* (0.073)	[0.631, 0.917]	1.920* (1.609)	[-1.234, 5.074]	-0.189* (1.956)	[-4.023, 3.644]	0.032* (2.256)	[-4.390, 4.453]	0.926* (4.668)	[-8.223, 10.076]
b44*EB Status	4.670* (0.075)	[4.523, 4.817]	5.744* (3.215)	[-0.557, 12.045]	3.705* (2.838)	[-1.857, 9.268]	3.931* (0.092)	[3.751, 4.112]	4.744* (4.406)	[-3.892, 13.380]
b45*EB Status	4.708* (0.077)	[4.557, 4.859]	5.878* (1.910)	[2.134, 9.621]	3.749* (2.458)	[-1.069, 8.567]	3.968* (0.100)	[3.772, 4.164]	4.899* (2.908)	[-0.800, 10.599]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G18.

EPvSTEB Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	-0.998*** (0.021)	-	21.175*** (1.033)	-	15.829*** (1.224)	-	-8.507*** (0.434)	-	13.473*** (1.655)	-
Intercept*STEB	1.280*** (0.066)	-	-14.535*** (1.595)	-	-8.746*** (2.185)	-	5.688*** (0.798)	-	-10.259** (3.251)	-
LEX	-	-	13.682*** (0.639)	-	-	-	-	-	8.551*** (0.794)	-
LEX*STEB	-	-	-9.719*** (0.989)	-	-	-	-	-	-6.793*** (1.386)	-
NP	-	-	-	-	20.609*** (1.500)	-	-	-	5.214*** (1.517)	-
NP*STEB	-	-	-	-	-12.241*** (2.683)	-	-	-	-3.164 (2.837)	-
RC	-	-	-	-	-	-	-28.809*** (1.664)	-	-14.07*** (1.758)	-
RC*STEB	-	-	-	-	-	-	17.072*** (3.021)	-	8.217* (3.734)	-
b01	-0.369*** (0.027)	-	-53.037*** (2.452)	-	-23.136*** (1.656)	-	72.616*** (4.212)	-	-3.408 (6.439)	-
b01*STEB	0.273** (0.084)	1.585	37.747*** (3.799)	2.810	13.804*** (2.964)	0.602	-43.002*** (7.649)	0.807	9.127 (13.433)	1.790
b02	-	-	-	-	-	-	-	-	-	-
b02*STEB	-	-	-	-	-	-	-	-	-	-
b03	0.344*** (0.026)	-	-24.510*** (1.158)	-	-21.553*** (1.593)	-	-1.960*** (0.136)	-	-21.888*** (1.618)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b03*STEB	-0.291*** (0.084)	1.009	17.391*** (1.795)	2.141	12.723*** (2.851)	0.024	1.074*** (0.256)	0.280	16.125*** (2.985)	1.216
b04	0.079** (0.026)	-	-27.712*** (1.295)	-	-8.437*** (0.620)	-	-2.226*** (0.136)	-	-20.610*** (1.406)	-
b04*STEB	-0.008 (0.084)	1.298	19.766*** (2.007)	2.442	5.053*** (1.111)	0.304	1.359*** (0.256)	0.571	15.822*** (2.431)	1.519
b05	0.478*** (0.026)	-	-24.374*** (1.158)	-	-4.783*** (0.383)	-	-1.825*** (0.136)	-	-17.538*** (1.266)	-
b05*STEB	-0.262** (0.085)	1.039	17.42*** (1.795)	2.171	2.865*** (0.689)	0.045	1.103*** (0.256)	0.310	13.595*** (2.165)	1.240
b06	-0.076** (0.027)	-	-10.317*** (0.478)	-	-23.172*** (1.680)	-	-2.383*** (0.136)	-	-13.468*** (1.532)	-
b06*STEB	-0.292*** (0.083)	1.008	7.000*** (0.744)	2.110	13.433*** (3.007)	0.024	1.073*** (0.256)	0.279	9.033** (2.883)	1.181
b07	0.060* (0.026)	-	-11.635*** (0.545)	-	-6.394*** (0.470)	-	-2.246*** (0.136)	-	-10.027*** (0.592)	-
b07*STEB	-0.621*** (0.083)	0.673	7.708*** (0.848)	1.781	3.217*** (0.844)	0.321	0.743** (0.256)	0.058	6.870*** (1.024)	0.845
b08	-0.952*** (0.028)	-	-30.211*** (1.362)	-	-3.592*** (0.194)	-	-2.999*** (0.121)	-	-20.957*** (1.569)	-
b08*STEB	0.276*** (0.084)	1.588	21.087*** (2.111)	2.739	1.848*** (0.354)	0.593	1.487*** (0.230)	0.859	15.855*** (2.689)	1.814
b09	-0.005 (0.026)	-	-43.880*** (2.044)	-	-13.450*** (0.978)	-	22.764*** (1.314)	-	-19.747*** (3.291)	-
b09*STEB	0.169* (0.085)	1.479	31.395*** (3.166)	2.676	8.159*** (1.752)	0.488	-13.332*** (2.388)	0.736	17.600** (6.502)	1.726
b10	0.91*** (0.026)	-	-42.902*** (2.044)	-	-2.954*** (0.282)	-	23.930*** (1.329)	-	-16.227*** (3.367)	-
b10*STEB	-0.121 (0.088)	1.183	31.077*** (3.167)	2.351	2.173*** (0.509)	0.188	-13.772*** (2.416)	0.444	15.737* (6.556)	1.401

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b11	-0.296*** (0.027)	-	-4.677*** (0.206)	-	-8.688*** (0.611)	-	-2.603*** (0.136)	-	-6.291*** (0.545)	-
b11*STEB	0.440*** (0.085)	1.755	3.550*** (0.327)	2.824	5.427*** (1.095)	0.760	1.809*** (0.256)	1.030	4.569*** (1.013)	1.888
b12	-3.938*** (0.026)	-	-47.715*** (2.044)	-	-32.771*** (2.099)	-	-6.233*** (0.136)	-	-39.697*** (2.510)	-
b12*STEB	4.364*** (0.087)	5.760	35.480*** (3.167)	6.845	21.494*** (3.755)	4.765	5.722*** (0.257)	5.024	31.189*** (4.573)	5.886
b13	-0.303*** (0.027)	-	-13.464*** (0.614)	-	-2.941*** (0.194)	-	-2.610*** (0.136)	-	-10.343*** (0.646)	-
b13*STEB	0.299*** (0.084)	1.612	9.661*** (0.953)	2.713	1.866*** (0.354)	0.612	1.667*** (0.256)	0.886	7.927*** (1.076)	1.779
b14	-0.268*** (0.027)	-	-26.598*** (1.226)	-	-15.692*** (1.122)	-	23.743*** (1.386)	-	-8.915*** (2.582)	-
b14*STEB	-0.071 (0.083)	1.234	18.66*** (1.901)	2.375	9.096*** (2.009)	0.245	-14.306*** (2.518)	0.491	8.552 (5.311)	1.416
b15	0.012 (0.026)	-	-16.071*** (0.750)	-	-0.606*** (0.052)	-	-2.293*** (0.136)	-	-11.345*** (0.856)	-
b15*STEB	0.050 (0.084)	1.357	11.492*** (1.163)	2.469	0.418*** (0.116)	0.358	1.417*** (0.256)	0.630	8.825*** (1.440)	1.535
b16	0.163*** (0.026)	-	-30.547*** (1.431)	-	-22.148*** (1.623)	-	-2.142*** (0.136)	-	-25.844*** (1.812)	-
b16*STEB	-0.002 (0.085)	1.304	21.850*** (2.217)	2.456	13.259*** (2.905)	0.321	1.365*** (0.256)	0.577	19.409*** (3.308)	1.543
b17	1.337*** (0.026)	-	-40.981*** (1.975)	-	-67.189*** (4.989)	-	71.928*** (4.077)	-	-8.012 (7.255)	-
b17*STEB	-0.625*** (0.088)	0.668	29.494*** (3.061)	1.797	40.101*** (8.926)	0.319	-42.476*** (7.405)	0.067	10.831 (15.013)	0.854
b18	0.114*** (0.026)	-	-7.197*** (0.341)	-	-17.249*** (1.263)	-	-2.191*** (0.136)	-	-9.991*** (1.149)	-

Effect	Base model Estimate (SE)	Base model Effect Size	LEX Predictor Estimate (SE)	LEX Predictor Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size	RC Predictor Estimate (SE)	RC Predictor Effect Size	All Predictors Estimate (SE)	All Predictors Effect Size
b18*STEB	-0.059 (0.084)	1.246	5.142*** (0.534)	2.336	10.260*** (2.261)	0.258	1.308*** (0.256)	0.519	6.919** (2.158)	1.405
b19	0.351*** (0.026)	-	0.334*** (0.025)	-	-0.271*** (0.052)	-	-1.953*** (0.136)	-	-0.953*** (0.147)	-
b19*STEB	-0.566*** (0.083)	0.729	-0.542*** (0.082)	1.822	-0.193+ (0.116)	0.265	0.798** (0.256)	0.001	0.217 (0.313)	0.882
b20	0.518*** (0.026)	-	-8.272*** (0.409)	-	-2.744*** (0.238)	-	-1.785*** (0.136)	-	-6.946*** (0.410)	-
b20*STEB	-0.472*** (0.084)	0.825	5.784*** (0.638)	1.930	1.468*** (0.432)	0.169	0.892*** (0.256)	0.095	5.079*** (0.675)	0.994
b21	0.680*** (0.026)	-	-13.96*** (0.682)	-	-70.186*** (5.157)	-	-1.622*** (0.136)	-	-27.547*** (4.919)	-
b21*STEB	-0.645*** (0.084)	0.648	9.773*** (1.059)	1.766	41.462*** (9.226)	0.329	0.718** (0.256)	0.083	18.206* (9.263)	0.840
b22	0.888*** (0.026)	-	-7.916*** (0.409)	-	0.259*** (0.052)	-	-1.413*** (0.136)	-	-5.922*** (0.441)	-
b22*STEB	-0.467*** (0.086)	0.830	5.794*** (0.639)	1.940	-0.093 (0.117)	0.163	0.897*** (0.256)	0.100	4.683*** (0.721)	1.003
b23	-3.317*** (0.027)	-	-35.458*** (1.498)	-	-20.786*** (1.271)	-	-4.518*** (0.075)	-	-28.434*** (1.754)	-
b23*STEB	3.284*** (0.087)	4.658	26.137*** (2.322)	5.780	13.669*** (2.274)	3.675	3.994*** (0.153)	3.923	22.302*** (3.164)	4.838
b24	-0.228*** (0.027)	-	-30.947*** (1.431)	-	-35.780*** (2.586)	-	-1.439*** (0.075)	-	-29.069*** (2.479)	-
b24*STEB	0.066 (0.084)	1.374	21.920*** (2.217)	2.528	21.195*** (4.627)	0.400	0.783*** (0.152)	0.645	21.203*** (4.669)	1.619
b25	-0.475*** (0.027)	-	-13.636*** (0.614)	-	-8.227*** (0.565)	-	-2.782*** (0.136)	-	-11.811*** (0.685)	-
b25*STEB	0.759*** (0.085)	2.081	10.117*** (0.953)	3.178	5.365*** (1.012)	1.084	2.129*** (0.256)	1.357	9.173*** (1.200)	2.250

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b26	0.464*** (0.026)	-	-5.408*** (0.274)	-	-0.160** (0.052)	-	-1.840*** (0.136)	-	-4.506*** (0.282)	-
b26*STEB	-0.303*** (0.084)	0.997	3.874*** (0.430)	2.094	0.068 (0.116)	0.001	1.062*** (0.256)	0.268	3.385*** (0.452)	1.155
b27	-0.377*** (0.027)	-	-42.801*** (1.975)	-	-8.253*** (0.574)	-	-2.137*** (0.105)	-	-29.811*** (2.234)	-
b27*STEB	-0.090 (0.083)	1.215	30.082*** (3.060)	2.397	4.592*** (1.028)	0.220	0.951*** (0.202)	0.486	22.769*** (3.877)	1.478
b28	-0.672*** (0.027)	-	-22.617*** (1.022)	-	-46.436*** (3.328)	-	95.202*** (5.533)	-	20.924** (7.280)	-
b28*STEB	1.035*** (0.086)	2.363	16.651*** (1.585)	3.499	28.242*** (5.955)	1.408	-55.827*** (10.048)	1.552	-8.490 (15.392)	2.398
b29	-0.492*** (0.027)	-	-6.339*** (0.274)	-	-8.863*** (0.610)	-	-2.799*** (0.136)	-	-7.400*** (0.541)	-
b29*STEB	-0.061 (0.083)	1.244	4.109*** (0.431)	2.334	4.916*** (1.093)	0.251	1.305*** (0.256)	0.516	4.812*** (0.999)	1.398
b30	0.198*** (0.026)	-	-23.205*** (1.090)	-	-8.174*** (0.609)	-	-2.107*** (0.136)	-	-17.705*** (1.175)	-
b30*STEB	-0.585*** (0.083)	0.709	16.063*** (1.691)	1.837	4.394*** (1.093)	0.281	0.779** (0.256)	0.021	13.025*** (2.034)	0.906
b31	-0.106*** (0.027)	-	-19.123*** (0.886)	-	-2.744*** (0.194)	-	-2.412*** (0.136)	-	-13.814*** (0.976)	-
b31*STEB	-0.100 (0.084)	1.204	13.431*** (1.374)	2.325	1.470*** (0.353)	0.207	1.267*** (0.256)	0.477	10.450*** (1.649)	1.394
b32	-3.304*** (0.027)	-	-31.097*** (1.295)	-	-24.478*** (1.541)	-	-5.598*** (0.136)	-	-27.180*** (1.672)	-
b32*STEB	3.209*** (0.087)	4.581	22.975*** (2.007)	5.717	15.793*** (2.757)	3.594	4.564*** (0.256)	3.842	20.945*** (3.058)	4.765
b33	1.249*** (0.026)	-	-7.569*** (0.409)	-	0.616*** (0.052)	-	-1.048*** (0.136)	-	-5.576*** (0.441)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b33*STEB	-0.492*** (0.088)	0.804	5.775*** (0.639)	1.921	-0.116 (0.119)	0.187	0.871*** (0.257)	0.073	4.664*** (0.721)	0.984
b34	-0.406*** (0.027)	-	-9.175*** (0.410)	-	-3.043*** (0.194)	-	-2.165*** (0.105)	-	-7.424*** (0.419)	-
b34*STEB	0.287*** (0.084)	1.599	6.527*** (0.639)	2.689	1.855*** (0.353)	0.600	1.330*** (0.203)	0.873	5.569*** (0.699)	1.751
b35	0.591*** (0.026)	-	-31.556*** (1.498)	-	-6.506*** (0.517)	-	-1.711*** (0.136)	-	-22.448*** (1.653)	-
b35*STEB	-0.388*** (0.085)	0.910	22.483*** (2.322)	2.051	3.83*** (0.927)	0.082	0.977*** (0.256)	0.181	17.382*** (2.848)	1.122
b36	-0.235*** (0.027)	-	-29.486*** (1.362)	-	-37.87*** (2.737)	-	73.008*** (4.227)	-	7.761 (5.941)	-
b36*STEB	0.565*** (0.085)	1.883	21.384*** (2.112)	3.042	22.938*** (4.898)	0.917	-42.873*** (7.677)	1.095	-0.064 (12.545)	1.992
b37	-0.335*** (0.027)	-	-17.886*** (0.818)	-	-9.449*** (0.663)	-	-1.547*** (0.075)	-	-14.231*** (0.932)	-
b37*STEB	0.067 (0.084)	1.375	12.555*** (1.269)	2.492	5.485*** (1.189)	0.382	0.785*** (0.152)	0.647	10.570*** (1.663)	1.563
b38	0.050+ (0.026)	-	-18.968*** (0.886)	-	-14.096*** (1.029)	-	0.050+ (0.026)	-	-15.443*** (1.175)	-
b38*STEB	0.021 (0.084)	1.328	13.551*** (1.374)	2.448	8.429*** (1.843)	0.339	0.021 (0.085)	0.599	11.681*** (2.183)	1.519
b39	-0.063* (0.027)	-	-7.369*** (0.341)	-	-0.062* (0.026)	-	-2.369*** (0.136)	-	-5.766*** (0.379)	-
b39*STEB	-0.534*** (0.083)	0.761	4.676*** (0.534)	1.861	-0.530*** (0.083)	0.235	0.831** (0.256)	0.032	3.780*** (0.619)	0.920
b40	-0.309*** (0.027)	-	-17.859*** (0.818)	-	-46.833*** (3.384)	-	-0.425*** (0.028)	-	-23.146*** (3.154)	-
b40*STEB	0.217* (0.084)	1.528	12.704*** (1.269)	2.645	27.869*** (6.054)	0.565	0.286*** (0.085)	0.800	16.168** (5.983)	1.737

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b41	0.279*** (0.026)	-	-14.354*** (0.682)	-	-4.877*** (0.376)	-	-1.478*** (0.105)	-	-11.051*** (0.726)	-
b41*STEB	-0.341*** (0.084)	0.958	10.072*** (1.059)	2.071	2.725*** (0.676)	0.036	0.700*** (0.202)	0.230	8.254*** (1.248)	1.137
b42	-0.536*** (0.027)	-	-38.572*** (1.771)	-	-72.821*** (5.257)	-	-2.843*** (0.136)	-	-43.801*** (4.836)	-
b42*STEB	0.585*** (0.085)	1.903	27.654*** (2.744)	3.093	43.552*** (9.406)	0.967	1.953*** (0.256)	1.177	31.363*** (9.163)	2.224
b43	0.355*** (0.026)	-	-23.047*** (1.090)	-	-15.667*** (1.166)	-	23.669*** (1.346)	-	-6.955** (2.445)	-
b43*STEB	-0.604*** (0.083)	0.690	16.043*** (1.691)	1.817	8.918*** (2.087)	0.299	-14.425*** (2.445)	0.048	6.856 (5.053)	0.867
b44	-3.436*** (0.027)	-	-50.111*** (2.18)	-	-26.689*** (1.692)	-	-2.880*** (0.042)	-	-38.201*** (2.581)	-
b44*STEB	3.504*** (0.087)	4.883	36.664*** (3.378)	5.940	17.322*** (3.028)	3.893	3.174*** (0.104)	4.148	30.080*** (4.696)	4.983
b45	-5.115*** (0.026)	-	-32.933*** (1.295)	-	-25.262*** (1.466)	-	-5.772*** (0.046)	-	-27.981*** (1.678)	-
b45*STEB	3.242*** (0.089)	4.615	23.024*** (2.007)	5.767	15.217*** (2.623)	3.631	3.627*** (0.113)	3.879	20.385*** (3.096)	4.833
delta1	5.361*** (0.010)	-	5.389*** (0.010)	-	5.365*** (0.010)	-	5.361*** (0.010)	-	5.409*** (0.010)	-
delta1*STEB	-1.121*** (0.034)	-	-1.130*** (0.035)	-	-1.126*** (0.034)	-	-1.114*** (0.034)	-	-1.124*** (0.035)	-
delta2	6.981*** (0.014)	-	7.053*** (0.014)	-	6.989*** (0.014)	-	6.988*** (0.014)	-	7.108*** (0.014)	-
delta2*STEB	-1.018*** (0.074)	-	-1.060*** (0.074)	-	-1.026*** (0.074)	-	-1.016*** (0.074)	-	-1.078*** (0.074)	-
delta3	8.324*** (0.019)	-	8.420*** (0.019)	-	8.335*** (0.019)	-	8.336*** (0.019)	-	8.497*** (0.019)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
delta3*STEB	-1.177*** (0.130)	-	-1.238*** (0.131)	-	-1.187*** (0.130)	-	-1.178*** (0.130)	-	-1.272*** (0.131)	-
Intercept Variance	1.048		1.038		1.037		1.028		1.042	
LEX Variance	-		0.03		-		-		0.052	
NP Variance	-		-		0.006		-		0.005	
RC Variance	-		-		-		0.006		0.040	
Intercept*Feature Covariance	-		0.173		0.068		-0.078		See Table G30	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G19.

EPvSTEB Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b01*STEB	1.553* (0.084)	[1.388, 1.718]	2.754* (3.799)	[-4.693, 10.200]	0.590* (2.964)	[-5.219, 6.399]	0.791 (7.649)	[-14.201, 15.783]	1.754 (13.433)	[-24.574, 28.083]
b02*STEB	-	-	-	-	-	-	-	-	-	-
b03*STEB	0.989* (0.084)	[0.824, 1.154]	2.098* (1.795)	[-1.420, 5.616]	0.023* (2.851)	[-5.565, 5.611]	0.275* (0.256)	[-0.227, 0.776]	1.192* (2.985)	[-4.659, 7.042]
b04*STEB	1.272 (0.084)	[1.107, 1.437]	2.393* (2.007)	[-1.541, 6.327]	0.298* (1.111)	[-1.880, 2.475]	0.560* (0.256)	[0.058, 1.061]	1.488* (2.431)	[-3.276, 6.253]
b05*STEB	1.018* (0.085)	[0.851, 1.185]	2.127* (1.795)	[-1.391, 5.645]	0.044* (0.689)	[-1.307, 1.394]	0.304* (0.256)	[-0.198, 0.805]	1.215* (2.165)	[-3.028, 5.458]
b06*STEB	0.988* (0.083)	[0.825, 1.151]	2.067* (0.744)	[0.609, 3.526]	0.023* (3.007)	[-5.871, 5.917]	0.274* (0.256)	[-0.228, 0.775]	1.158* (2.883)	[-4.493, 6.808]
b07*STEB	0.659* (0.083)	[0.496, 0.822]	1.745* (0.848)	[0.083, 3.407]	-0.314* (0.844)	[-1.969, 1.340]	-0.056* (0.256)	[-0.558, 0.445]	0.828* (1.024)	[-1.179, 2.835]
b08*STEB	1.556* (0.084)	[1.391, 1.721]	2.684* (2.111)	[-1.454, 6.821]	0.581* (0.354)	[-0.113, 1.275]	0.841* (0.230)	[0.390, 1.292]	1.777* (2.689)	[-3.493, 7.048]
b09*STEB	1.449* (0.085)	[1.282, 1.616]	2.622* (3.166)	[-3.584, 8.827]	0.478* (1.752)	[-2.956, 3.912]	0.721* (2.388)	[-3.959, 5.402]	1.691 (6.502)	[-11.053, 14.435]
b10*STEB	1.159 (0.088)	[0.987, 1.331]	2.304* (3.167)	[-3.904, 8.511]	0.184* (0.509)	[-0.814, 1.182]	0.435* (2.416)	[-4.300, 5.170]	1.373 (6.556)	[-11.477, 14.223]
b11*STEB	1.720* (0.085)	[1.553, 1.887]	2.767* (0.327)	[2.126, 3.408]	0.745* (1.095)	[-1.401, 2.891]	1.010* (0.256)	[0.508, 1.511]	1.85* (1.013)	[-0.135, 3.836]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b12*STEB	5.644* (0.087)	[5.473, 5.815]	6.707* (3.167)	[0.499, 12.914]	4.669* (3.755)	[-2.691, 12.029]	4.923* (0.257)	[4.419, 5.426]	5.768* (4.573)	[-3.196, 14.731]
b13*STEB	1.579* (0.084)	[1.414, 1.744]	2.658* (0.953)	[0.790, 4.526]	0.599* (0.354)	[-0.095, 1.293]	0.868* (0.256)	[0.366, 1.369]	1.743* (1.076)	[-0.366, 3.852]
b14*STEB	1.209 (0.083)	[1.046, 1.372]	2.327* (1.901)	[-1.399, 6.053]	0.240* (2.009)	[-3.698, 4.177]	0.481* (2.518)	[-4.454, 5.417]	1.387* (5.311)	[-9.022, 11.797]
b15*STEB	1.330 (0.084)	[1.165, 1.495]	2.419* (1.163)	[0.140, 4.699]	0.351* (0.116)	[0.124, 0.578]	0.618* (0.256)	[0.116, 1.119]	1.504* (1.440)	[-1.318, 4.327]
b16*STEB	1.278 (0.085)	[1.111, 1.445]	2.407* (2.217)	[-1.938, 6.752]	0.314* (2.905)	[-5.379, 6.008]	0.566* (0.256)	[0.064, 1.067]	1.512* (3.308)	[-4.972, 7.996]
b17*STEB	0.655* (0.088)	[0.483, 0.827]	1.761* (3.061)	[-4.239, 7.760]	-0.312 (8.926)	[-17.807, 17.182]	-0.066 (7.405)	[-14.580, 14.448]	0.837 (15.013)	[-28.589, 30.262]
b18*STEB	1.221 (0.084)	[1.056, 1.386]	2.289* (0.534)	[1.243, 3.336]	0.253* (2.261)	[-4.178, 4.685]	0.509* (0.256)	[0.007, 1.010]	1.377* (2.158)	[-2.853, 5.607]
b19*STEB	0.714* (0.083)	[0.551, 0.877]	1.785* (0.082)	[1.625, 1.946]	-0.260* (0.116)	[-0.487, -0.033]	-0.001* (0.256)	[-0.503, 0.500]	0.865* (0.313)	[0.251, 1.478]
b20*STEB	0.808* (0.084)	[0.643, 0.973]	1.891* (0.638)	[0.641, 3.142]	-0.166* (0.432)	[-1.013, 0.681]	0.093* (0.256)	[-0.409, 0.594]	0.974* (0.675)	[-0.349, 2.297]
b21*STEB	0.635* (0.084)	[0.470, 0.800]	1.730* (1.059)	[-0.345, 3.806]	-0.322 (9.226)	[-18.405, 17.761]	-0.081* (0.256)	[-0.583, 0.420]	0.823 (9.263)	[-17.333, 18.978]
b22*STEB	0.813* (0.086)	[0.644, 0.982]	1.901* (0.639)	[0.649, 3.154]	-0.160* (0.117)	[-0.389, 0.069]	0.098* (0.256)	[-0.404, 0.599]	0.983* (0.721)	[-0.430, 2.396]
b23*STEB	4.564* (0.087)	[4.393, 4.735]	5.664* (2.322)	[1.113, 10.215]	3.601* (2.274)	[-0.856, 8.058]	3.843* (0.153)	[3.543, 4.143]	4.741* (3.164)	[-1.461, 10.942]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
b24*STEB	1.346 (0.084)	[1.181, 1.511]	2.477* (2.217)	[-1.868, 6.822]	0.392* (4.627)	[-8.677, 9.461]	0.632* (0.152)	[0.334, 0.930]	1.587* (4.669)	[-7.564, 10.738]
b25*STEB	2.039* (0.085)	[1.872, 2.206]	3.114* (0.953)	[1.246, 4.982]	1.062* (1.012)	[-0.921, 3.046]	1.330* (0.256)	[0.828, 1.831]	2.205* (1.200)	[-0.147, 4.557]
b26*STEB	0.977* (0.084)	[0.812, 1.142]	2.051* (0.430)	[1.209, 2.894]	0.001* (0.116)	[-0.226, 0.228]	0.263* (0.256)	[-0.239, 0.764]	1.132* (0.452)	[0.246, 2.018]
b27*STEB	1.190 (0.083)	[1.027, 1.353]	2.349* (3.060)	[-3.649, 8.346]	0.216* (1.028)	[-1.799, 2.231]	0.476* (0.202)	[0.080, 0.872]	1.448* (3.877)	[-6.151, 9.047]
b28*STEB	2.315* (0.086)	[2.146, 2.484]	3.428* (1.585)	[0.321, 6.535]	1.379 (5.955)	[-10.292, 13.051]	1.521 (10.048)	[-18.173, 21.215]	2.350 (15.392)	[-27.818, 32.518]
b29*STEB	1.219 (0.083)	[1.056, 1.382]	2.286* (0.431)	[1.442, 3.131]	0.246* (1.093)	[-1.896, 2.389]	0.506* (0.256)	[0.004, 1.007]	1.369* (0.999)	[-0.589, 3.327]
b30*STEB	0.695* (0.083)	[0.532, 0.858]	1.800* (1.691)	[-1.514, 5.114]	-0.276* (1.093)	[-2.418, 1.867]	-0.020* (0.256)	[-0.522, 0.481]	0.887* (2.034)	[-3.099, 4.874]
b31*STEB	1.180 (0.084)	[1.015, 1.345]	2.278* (1.374)	[-0.415, 4.971]	0.203* (0.353)	[-0.489, 0.895]	0.468* (0.256)	[-0.034, 0.969]	1.366* (1.649)	[-1.866, 4.598]
b32*STEB	4.489* (0.087)	[4.318, 4.660]	5.602* (2.007)	[1.668, 9.536]	3.522* (2.757)	[-1.882, 8.925]	3.765* (0.256)	[3.263, 4.266]	4.669* (3.058)	[-1.325, 10.662]
b33*STEB	0.788* (0.088)	[0.616, 0.960]	1.882* (0.639)	[0.630, 3.135]	-0.183* (0.119)	[-0.416, 0.050]	0.072* (0.257)	[-0.432, 0.575]	0.964* (0.721)	[-0.449, 2.377]
b34*STEB	1.567* (0.084)	[1.402, 1.732]	2.634* (0.639)	[1.382, 3.887]	0.588* (0.353)	[-0.104, 1.280]	0.855* (0.203)	[0.457, 1.253]	1.715* (0.699)	[0.345, 3.085]
b35*STEB	0.892* (0.085)	[0.725, 1.059]	2.010* (2.322)	[-2.541, 6.561]	-0.081* (0.927)	[-1.898, 1.736]	0.178* (0.256)	[-0.324, 0.679]	1.100* (2.848)	[-4.482, 6.682]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b36*STEB	1.845* (0.085)	[1.678, 2.012]	2.981* (2.112)	[-1.159, 7.120]	0.898* (4.898)	[-8.702, 10.498]	1.073 (7.677)	[-13.974, 16.120]	1.952 (12.545)	[-22.637, 26.540]
b37*STEB	1.347 (0.084)	[1.182, 1.512]	2.442* (1.269)	[-0.045, 4.929]	0.375* (1.189)	[-1.956, 2.705]	0.634* (0.152)	[0.336, 0.932]	1.531* (1.663)	[-1.728, 4.791]
b38*STEB	1.301 (0.084)	[1.136, 1.466]	2.398* (1.374)	[-0.295, 5.091]	0.332* (1.843)	[-3.281, 3.944]	0.587* (0.085)	[0.421, 0.754]	1.489* (2.183)	[-2.790, 5.767]
b39*STEB	0.746* (0.083)	[0.583, 0.909]	1.823* (0.534)	[0.777, 2.870]	-0.230* (0.083)	[-0.393, -0.067]	0.032* (0.256)	[-0.470, 0.533]	0.902* (0.619)	[-0.311, 2.115]
b40*STEB	1.497* (0.084)	[1.332, 1.662]	2.591* (1.269)	[0.104, 5.078]	0.553 (6.054)	[-11.312, 12.419]	0.784* (0.085)	[0.618, 0.951]	1.702* (5.983)	[-10.025, 13.429]
b41*STEB	0.939* (0.084)	[0.774, 1.104]	2.029* (1.059)	[-0.046, 4.105]	-0.035* (0.676)	[-1.360, 1.290]	0.225* (0.202)	[-0.171, 0.621]	1.114* (1.248)	[-1.332, 3.560]
b42*STEB	1.865* (0.085)	[1.698, 2.032]	3.031* (2.744)	[-2.348, 8.409]	0.947 (9.406)	[-17.488, 19.383]	1.154* (0.256)	[0.652, 1.655]	2.179 (9.163)	[-15.781, 20.138]
b43*STEB	0.676* (0.083)	[0.513, 0.839]	1.780* (1.691)	[-1.534, 5.094]	-0.293* (2.087)	[-4.384, 3.797]	-0.047* (2.445)	[-4.840, 4.745]	0.849* (5.053)	[-9.054, 10.753]
b44*STEB	4.784* (0.087)	[4.613, 4.955]	5.821* (3.378)	[-0.800, 12.441]	3.814* (3.028)	[-2.121, 9.749]	4.065* (0.104)	[3.861, 4.269]	4.883* (4.696)	[-4.322, 14.087]
b45*STEB	4.522* (0.089)	[4.348, 4.696]	5.651* (2.007)	[1.717, 9.585]	3.558* (2.623)	[-1.583, 8.699]	3.801* (0.113)	[3.579, 4.022]	4.735* (3.096)	[-1.333, 10.803]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G20.

EPvLTEB Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
Intercept	-0.998*** (0.021)	-	21.182*** (1.042)	-	15.806*** (1.234)	-	-8.509*** (0.438)	-	13.437*** (1.669)	-
Intercept*LTEB	1.386*** (0.107)	-	-15.531*** (3.368)	-	-7.086 (4.568)	-	6.585*** (1.484)	-	-14.536* (6.872)	-
LEX	-	-	13.687*** (0.644)	-	-	-	-	-	8.544*** (0.801)	-
LEX*LTEB	-	-	-10.417*** (2.085)	-	-	-	-	-	-7.558* (2.987)	-
NP	-	-	-	-	20.581*** (1.512)	-	-	-	5.185*** (1.529)	-
NP*LTEB	-	-	-	-	-10.350+ (5.604)	-	-	-	-6.546 (6.449)	-
RC	-	-	-	-	-	-	-28.812*** (1.679)	-	-14.070*** (1.772)	-
RC*LTEB	-	-	-	-	-	-	19.993*** (5.649)	-	6.619 (7.061)	-
b01	-0.370*** (0.027)	-	-53.054*** (2.474)	-	-23.106*** (1.669)	-	72.624*** (4.250)	-	-3.345 (6.492)	-
b01*LTEB	0.238+ (0.135)	1.657	40.394*** (8.009)	2.996	11.683+ (6.188)	0.836	-50.434*** (14.303)	0.791	19.811 (26.524)	1.786
b02	-	-	-	-	-	-	-	-	-	-
b02*LTEB	-	-	-	-	-	-	-	-	-	-
b03	0.344*** (0.026)	-	-24.518*** (1.168)	-	-21.524*** (1.606)	-	-1.960*** (0.137)	-	-21.844*** (1.631)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b03*LTEB	-0.398** (0.135)	1.008	18.550*** (3.783)	2.252	10.608+ (5.953)	0.183	1.201* (0.472)	0.193	20.871** (6.465)	1.139
b04	0.079** (0.026)	-	-27.721*** (1.306)	-	-8.426*** (0.625)	-	-2.226*** (0.137)	-	-20.583*** (1.417)	-
b04*LTEB	-0.269* (0.134)	1.140	20.917*** (4.229)	2.393	4.013+ (2.318)	0.307	1.331** (0.472)	0.325	18.377*** (5.182)	1.279
b05	0.478*** (0.026)	-	-24.382*** (1.168)	-	-4.776*** (0.386)	-	-1.825*** (0.137)	-	-17.517*** (1.276)	-
b05*LTEB	-0.457*** (0.135)	0.948	18.489*** (3.783)	2.190	2.189 (1.435)	0.115	1.141* (0.472)	0.131	15.520*** (4.646)	1.069
b06	-0.077** (0.027)	-	-10.320*** (0.482)	-	-23.142*** (1.694)	-	-2.383*** (0.137)	-	-13.431*** (1.544)	-
b06*LTEB	-0.408** (0.134)	0.998	7.405*** (1.564)	2.211	11.201+ (6.278)	0.175	1.192* (0.471)	0.183	13.151* (6.437)	1.095
b07	0.060* (0.026)	-	-11.639*** (0.550)	-	-6.386*** (0.474)	-	-2.246*** (0.137)	-	-10.012*** (0.597)	-
b07*LTEB	-0.406** (0.134)	1.000	8.515*** (1.784)	2.217	2.839 (1.759)	0.165	1.193* (0.471)	0.184	8.667*** (2.183)	1.093
b08	-0.953*** (0.028)	-	-30.221*** (1.374)	-	-3.589*** (0.196)	-	-3.000*** (0.122)	-	-20.938*** (1.581)	-
b08*LTEB	0.111 (0.134)	1.528	22.413*** (4.450)	2.792	1.441* (0.730)	0.693	1.530*** (0.423)	0.712	17.645** (5.811)	1.679
b09	-0.005 (0.026)	-	-43.895*** (2.062)	-	-13.432*** (0.986)	-	22.767*** (1.326)	-	-19.704*** (3.318)	-
b09*LTEB	0.172 (0.136)	1.590	33.631*** (6.675)	2.898	6.930+ (3.656)	0.760	-15.638*** (4.465)	0.759	23.520+ (13.162)	1.760
b10	0.910*** (0.026)	-	-42.917*** (2.062)	-	-2.949*** (0.284)	-	23.934*** (1.341)	-	-16.197*** (3.395)	-
b10*LTEB	-0.335* (0.141)	1.073	33.078*** (6.675)	2.333	1.609 (1.057)	0.241	-16.318*** (4.516)	0.249	19.867 (13.211)	1.199

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b11	-0.296*** (0.027)	-	-4.679*** (0.208)	-	-8.677*** (0.616)	-	-2.603*** (0.137)	-	-6.278*** (0.549)	-
b11*LTEB	0.459*** (0.136)	1.883	3.793*** (0.681)	3.064	4.677* (2.285)	1.048	2.062*** (0.472)	1.071	6.082** (2.220)	1.938
b12	-3.940*** (0.026)	-	-47.732*** (2.062)	-	-32.734*** (2.116)	-	-6.235*** (0.137)	-	-39.636*** (2.531)	-
b12*LTEB	3.721*** (0.135)	5.212	37.051*** (6.676)	6.388	18.203* (7.841)	4.374	5.310*** (0.471)	4.386	37.564*** (9.781)	5.225
b13	-0.304*** (0.027)	-	-13.469*** (0.619)	-	-2.938*** (0.195)	-	-2.611*** (0.137)	-	-10.332*** (0.651)	-
b13*LTEB	0.191 (0.135)	1.609	10.225*** (2.006)	2.824	1.517* (0.730)	0.770	1.793*** (0.472)	0.797	8.861*** (2.332)	1.701
b14	-0.268*** (0.027)	-	-26.607*** (1.237)	-	-15.672*** (1.131)	-	23.745*** (1.398)	-	-8.878*** (2.603)	-
b14*LTEB	-0.258* (0.134)	1.151	19.814*** (4.006)	2.404	7.495+ (4.194)	0.322	-16.927*** (4.707)	0.321	13.696 (10.750)	1.256
b15	0.012 (0.026)	-	-16.076*** (0.756)	-	-0.605*** (0.052)	-	-2.293*** (0.137)	-	-11.336*** (0.863)	-
b15*LTEB	-0.125 (0.135)	1.287	12.138*** (2.450)	2.512	0.186 (0.215)	0.447	1.475** (0.472)	0.472	9.522** (3.160)	1.387
b16	0.163*** (0.026)	-	-30.557*** (1.443)	-	-22.118*** (1.636)	-	-2.142*** (0.137)	-	-25.796*** (1.827)	-
b16*LTEB	0.089 (0.137)	1.505	23.506*** (4.673)	2.770	11.304+ (6.065)	0.682	1.690*** (0.472)	0.692	24.750*** (7.105)	1.670
b17	1.338*** (0.026)	-	-40.994*** (1.993)	-	-67.096*** (5.029)	-	71.937*** (4.114)	-	-7.890 (7.315)	-
b17*LTEB	-0.755*** (0.141)	0.644	31.510*** (6.452)	1.870	33.675+ (18.639)	0.190	-49.760*** (13.846)	0.174	28.231 (31.741)	0.749
b18	0.114*** (0.026)	-	-7.200*** (0.344)	-	-17.226*** (1.273)	-	-2.192*** (0.137)	-	-9.963*** (1.159)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b18*LTEB	-0.155 (0.135)	1.256	5.420*** (1.120)	2.460	8.572+ (4.720)	0.429	1.445** (0.472)	0.442	9.953* (4.811)	1.339
b19	0.351*** (0.026)	-	0.334*** (0.025)	-	-0.270*** (0.052)	-	-1.953*** (0.137)	-	-0.952*** (0.149)	-
b19*LTEB	-0.095 (0.137)	1.318	-0.085 (0.135)	2.508	0.218 (0.217)	0.480	1.505** (0.472)	0.503	0.645 (0.573)	1.376
b20	0.518*** (0.026)	-	-8.275*** (0.413)	-	-2.740*** (0.240)	-	-1.785*** (0.137)	-	-6.936*** (0.413)	-
b20*LTEB	-0.417** (0.136)	0.989	6.286*** (1.341)	2.206	1.225 (0.896)	0.155	1.181* (0.472)	0.172	6.030*** (1.444)	1.080
b21	0.681*** (0.026)	-	-13.964*** (0.688)	-	-70.091*** (5.198)	-	-1.621*** (0.137)	-	-27.439*** (4.960)	-
b21*LTEB	-0.682*** (0.135)	0.718	10.481*** (2.229)	1.948	34.926+ (19.267)	0.097	0.914+ (0.472)	0.100	30.487 (20.924)	0.834
b22	0.888*** (0.026)	-	-7.918*** (0.413)	-	0.260*** (0.052)	-	-1.412*** (0.137)	-	-5.916*** (0.445)	-
b22*LTEB	-0.698*** (0.137)	0.702	6.017*** (1.341)	1.932	-0.378+ (0.216)	0.128	0.898+ (0.472)	0.117	4.923** (1.599)	0.805
b23	-3.318*** (0.027)	-	-35.470*** (1.512)	-	-20.764*** (1.281)	-	-4.519*** (0.076)	-	-28.395*** (1.768)	-
b23*LTEB	3.170*** (0.136)	4.650	27.664*** (4.894)	5.887	11.954* (4.748)	3.827	4.001*** (0.273)	3.826	26.788*** (6.729)	4.759
b24	-0.228*** (0.027)	-	-30.957*** (1.443)	-	-35.734*** (2.607)	-	-1.440*** (0.075)	-	-29.003*** (2.499)	-
b24*LTEB	0.103 (0.135)	1.520	23.524*** (4.673)	2.789	17.975+ (9.662)	0.709	0.944*** (0.273)	0.706	28.725** (10.244)	1.695
b25	-0.475*** (0.027)	-	-13.640*** (0.619)	-	-8.217*** (0.569)	-	-2.783*** (0.137)	-	-11.793*** (0.691)	-
b25*LTEB	0.275* (0.134)	1.695	10.310*** (2.007)	2.911	4.172* (2.111)	0.860	1.877*** (0.472)	0.882	10.573*** (2.560)	1.792

Effect	Base model Estimate (SE)	Base model Effect Size	LEX Predictor Estimate (SE)	LEX Predictor Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size	RC Predictor Estimate (SE)	RC Predictor Effect Size	All Predictors Estimate (SE)	All Predictors Effect Size
b26	0.464*** (0.026)	-	-5.409*** (0.276)	-	-0.159** (0.052)	-	-1.839*** (0.137)	-	-4.501*** (0.284)	-
b26*LTEB	-0.047 (0.139)	1.367	4.424*** (0.901)	2.570	0.266 (0.218)	0.529	1.553** (0.473)	0.552	3.936*** (1.002)	1.441
b27	-0.378*** (0.027)	-	-42.815*** (1.993)	-	-8.243*** (0.578)	-	-2.137*** (0.106)	-	-29.777*** (2.252)	-
b27*LTEB	-0.158 (0.134)	1.253	32.175*** (6.452)	2.549	3.802+ (2.145)	0.419	1.062** (0.370)	0.438	26.266** (8.284)	1.443
b28	-0.673*** (0.028)	-	-22.624*** (1.031)	-	-46.376*** (3.355)	-	95.212*** (5.583)	-	21.002** (7.340)	-
b28*LTEB	0.508*** (0.135)	1.933	17.240*** (3.339)	3.179	23.514* (12.436)	1.133	-66.055*** (18.788)	1.050	5.059 (31.337)	1.923
b29	-0.492*** (0.027)	-	-6.341*** (0.276)	-	-8.852*** (0.614)	-	-2.799*** (0.137)	-	-7.386*** (0.546)	-
b29*LTEB	0.230+ (0.134)	1.649	4.691*** (0.900)	2.843	4.438+ (2.279)	0.815	1.831*** (0.472)	0.836	6.669** (2.179)	1.718
b30	0.198*** (0.026)	-	-23.213*** (1.100)	-	-8.163*** (0.614)	-	-2.107*** (0.137)	-	-17.68*** (1.185)	-
b30*LTEB	-0.470*** (0.134)	0.935	17.372*** (3.562)	2.177	3.739 (2.279)	0.102	1.129* (0.471)	0.119	15.696*** (4.333)	1.058
b31	-0.106*** (0.027)	-	-19.130*** (0.894)	-	-2.741*** (0.195)	-	-2.412*** (0.137)	-	-13.800*** (0.984)	-
b31*LTEB	-0.058 (0.135)	1.355	14.442*** (2.896)	2.588	1.269+ (0.730)	0.517	1.544** (0.471)	0.543	11.861*** (3.574)	1.469
b32	-3.305*** (0.027)	-	-31.108*** (1.306)	-	-24.452*** (1.553)	-	-5.599*** (0.137)	-	-27.137*** (1.686)	-
b32*LTEB	3.505*** (0.139)	4.992	24.694*** (4.229)	6.247	14.147* (5.756)	4.164	5.094*** (0.472)	4.166	26.158*** (6.583)	5.118
b33	1.250*** (0.026)	-	-7.571*** (0.413)	-	0.617*** (0.052)	-	-1.048*** (0.137)	-	-5.570*** (0.445)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b33*LTEB	-0.869*** (0.139)	0.528	5.856*** (1.342)	1.767	-0.546* (0.217)	0.300	0.725 (0.473)	0.293	4.763** (1.599)	0.642
b34	-0.406*** (0.027)	-	-9.178*** (0.413)	-	-3.040*** (0.195)	-	-2.166*** (0.106)	-	-7.416*** (0.423)	-
b34*LTEB	-0.012 (0.134)	1.402	6.680*** (1.341)	2.608	1.315+ (0.730)	0.564	1.209** (0.370)	0.588	6.100*** (1.502)	1.480
b35	0.591*** (0.026)	-	-31.566*** (1.512)	-	-6.496*** (0.521)	-	-1.711*** (0.137)	-	-22.421*** (1.667)	-
b35*LTEB	-0.401** (0.137)	1.005	24.108*** (4.894)	2.258	3.167 (1.932)	0.173	1.197* (0.472)	0.188	20.200*** (6.094)	1.140
b36	-0.235*** (0.027)	-	-29.496*** (1.374)	-	-37.821*** (2.760)	-	73.016*** (4.265)	-	7.831 (5.990)	-
b36*LTEB	0.340* (0.136)	1.762	22.645*** (4.450)	3.029	19.261+ (10.228)	0.954	-50.515*** (14.354)	0.892	11.585 (25.464)	1.801
b37	-0.336*** (0.027)	-	-17.892*** (0.825)	-	-9.437*** (0.669)	-	-1.547*** (0.076)	-	-14.209*** (0.939)	-
b37*LTEB	0.328* (0.136)	1.749	13.708*** (2.673)	2.977	4.908* (2.481)	0.914	1.169*** (0.273)	0.935	13.240*** (3.536)	1.861
b38	0.050+ (0.026)	-	-18.974*** (0.894)	-	-14.077*** (1.038)	-	0.050+ (0.026)	-	-15.413*** (1.185)	-
b38*LTEB	-0.093 (0.135)	1.320	14.406*** (2.896)	2.552	7.017+ (3.847)	0.489	-0.094 (0.135)	0.503	14.949** (4.685)	1.433
b39	-0.063* (0.027)	-	-7.372*** (0.344)	-	-0.062* (0.026)	-	-2.370*** (0.137)	-	-5.762*** (0.382)	-
b39*LTEB	-0.462*** (0.134)	0.943	5.117*** (1.120)	2.151	-0.460*** (0.133)	0.105	1.138* (0.471)	0.128	4.130** (1.390)	1.022
b40	-0.309*** (0.027)	-	-17.865*** (0.825)	-	-46.772*** (3.411)	-	-0.425*** (0.028)	-	-23.073*** (3.181)	-
b40*LTEB	0.199 (0.135)	1.618	13.581*** (2.673)	2.847	23.588+ (12.643)	0.818	0.280* (0.137)	0.803	24.758+ (13.457)	1.754

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b41	0.279*** (0.026)	-	-14.359*** (0.688)	-	-4.871*** (0.379)	-	-1.478*** (0.106)	-	-11.036*** (0.732)	-
b41*LTEB	-0.489*** (0.134)	0.915	10.670*** (2.229)	2.141	2.105 (1.407)	0.082	0.729* (0.370)	0.099	9.671*** (2.662)	1.016
b42	-0.537*** (0.027)	-	-38.585*** (1.787)	-	-72.727*** (5.300)	-	-2.844*** (0.137)	-	-43.682*** (4.877)	-
b42*LTEB	0.426** (0.135)	1.849	29.428*** (5.785)	3.148	36.766+ (19.642)	1.074	2.028*** (0.472)	1.037	45.047* (20.455)	2.086
b43	0.355*** (0.026)	-	-23.054*** (1.100)	-	-15.645*** (1.175)	-	23.672*** (1.358)	-	-6.919** (2.466)	-
b43*LTEB	-0.334* (0.136)	1.074	17.507*** (3.562)	2.314	7.720+ (4.356)	0.246	-16.520*** (4.572)	0.246	12.352 (10.273)	1.172
b44	-3.437*** (0.027)	-	-50.128*** (2.199)	-	-26.659*** (1.706)	-	-2.881*** (0.042)	-	-38.146*** (2.602)	-
b44*LTEB	3.014*** (0.136)	4.491	38.548*** (7.120)	5.651	14.698* (6.322)	3.660	2.626*** (0.173)	3.667	36.036*** (9.941)	4.502
b45	-5.116*** (0.026)	-	-32.944*** (1.306)	-	-25.237*** (1.477)	-	-5.774*** (0.047)	-	-27.940*** (1.692)	-
b45*LTEB	3.816*** (0.139)	5.309	25.015*** (4.229)	6.575	13.942* (5.476)	4.483	4.267*** (0.190)	4.485	25.782*** (6.643)	5.453
delta1	5.363*** (0.010)	-	5.392*** (0.010)	-	5.368*** (0.010)	-	5.364*** (0.010)	-	5.411*** (0.010)	-
delta1*LTEB	-1.168*** (0.058)	-	-1.184*** (0.058)	-	-1.172*** (0.058)	-	-1.167*** (0.058)	-	-1.189*** (0.059)	-
delta2	6.984*** (0.014)	-	7.056*** (0.014)	-	6.993*** (0.014)	-	6.991*** (0.014)	-	7.111*** (0.014)	-
delta2*LTEB	-0.562*** (0.169)	-	-0.616*** (0.169)	-	-0.569*** (0.169)	-	-0.566*** (0.169)	-	-0.650*** (0.170)	-
delta3	8.328*** (0.019)	-	8.423*** (0.019)	-	8.339*** (0.019)	-	8.340*** (0.019)	-	8.500*** (0.019)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
delta3*LTEB	-0.714* (0.305)	-	-0.789* (0.306)	-	-0.724* (0.305)	-	-0.722* (0.305)	-	-0.845** (0.306)	-
Intercept Variance	1.07		1.058		1.058		1.05		1.063	
LEX Variance	-		0.029		-		-		0.052	
NP Variance	-		-		0.006		-		0.006	
RC Variance	-		-		-		0.006		0.041	
Intercept*Feature Covariance	-		0.174		0.072		-0.079		See Table G30	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G21.

EPvLTEB Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b01*LTEB	1.624 (0.135)	[1.359, 1.889]	2.935* (8.009)	[-12.762, 18.633]	0.819 (6.188)	[-11.309, 12.948]	0.775 (14.303)	[-27.259, 28.809]	1.750 (26.524)	[-50.237, 53.737]
b02*LTEB	-	-	-	-	-	-	-	-	-	-
b03*LTEB	0.988* (0.135)	[0.723, 1.253]	2.206* (3.783)	[-5.208, 9.621]	0.179 (5.953)	[-11.489, 11.847]	0.189* (0.472)	[-0.736, 1.114]	1.116* (6.465)	[-11.556, 13.787]
b04*LTEB	1.117* (0.134)	[0.854, 1.380]	2.344* (4.229)	[-5.945, 10.633]	0.301* (2.318)	[-4.242, 4.844]	0.319* (0.472)	[-0.606, 1.244]	1.253* (5.182)	[-8.904, 11.410]
b05*LTEB	0.929* (0.135)	[0.664, 1.194]	2.145* (3.783)	[-5.269, 9.560]	0.112* (1.435)	[-2.700, 2.925]	0.129* (0.472)	[-0.796, 1.054]	1.048* (4.646)	[-8.059, 10.154]
b06*LTEB	0.978* (0.134)	[0.715, 1.241]	2.166* (1.564)	[-0.899, 5.231]	0.172 (6.278)	[-12.133, 12.477]	0.180* (0.471)	[-0.743, 1.103]	1.073* (6.437)	[-11.543, 13.690]
b07*LTEB	0.980* (0.134)	[0.717, 1.243]	2.172* (1.784)	[-1.325, 5.668]	0.162* (1.759)	[-3.286, 3.610]	0.181* (0.471)	[-0.742, 1.104]	1.071* (2.183)	[-3.208, 5.349]
b08*LTEB	1.497 (0.134)	[1.234, 1.760]	2.736* (4.450)	[-5.986, 11.458]	0.679* (0.730)	[-0.752, 2.110]	0.698* (0.423)	[-0.131, 1.527]	1.645* (5.811)	[-9.745, 13.034]
b09*LTEB	1.558 (0.136)	[1.291, 1.825]	2.839* (6.675)	[-10.244, 15.922]	0.744* (3.656)	[-6.421, 7.910]	0.744 (4.465)	[-8.008, 9.495]	1.724 (13.162)	[-24.073, 27.522]
b10*LTEB	1.051* (0.141)	[0.775, 1.327]	2.286* (6.675)	[-10.797, 15.369]	0.236* (1.057)	[-1.836, 2.308]	0.244 (4.516)	[-8.608, 9.095]	1.175 (13.211)	[-24.719, 27.068]
b11*LTEB	1.845* (0.136)	[1.578, 2.112]	3.002* (0.681)	[1.667, 4.337]	1.027* (2.285)	[-3.451, 5.506]	1.050* (0.472)	[0.125, 1.975]	1.899* (2.220)	[-2.453, 6.250]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b12*LTEB	5.107* (0.135)	[4.842, 5.372]	6.259* (6.676)	[-6.826, 19.344]	4.286 (7.841)	[-11.082, 19.654]	4.298* (0.471)	[3.375, 5.221]	5.120* (9.781)	[-14.051, 24.291]
b13*LTEB	1.577 (0.135)	[1.312, 1.842]	2.767* (2.006)	[-1.165, 6.699]	0.755* (0.730)	[-0.676, 2.186]	0.781* (0.472)	[-0.144, 1.706]	1.667* (2.332)	[-2.904, 6.238]
b14*LTEB	1.128 (0.134)	[0.865, 1.391]	2.356* (4.006)	[-5.496, 10.208]	0.316 (4.194)	[-7.904, 8.536]	0.314 (4.707)	[-8.911, 9.540]	1.231 (10.750)	[-19.839, 22.301]
b15*LTEB	1.261 (0.135)	[0.996, 1.526]	2.461* (2.450)	[-2.341, 7.263]	0.438* (0.215)	[0.017, 0.860]	0.463* (0.472)	[-0.462, 1.388]	1.359* (3.160)	[-4.834, 7.553]
b16*LTEB	1.475 (0.137)	[1.206, 1.744]	2.714* (4.673)	[-6.445, 11.873]	0.668 (6.065)	[-11.219, 12.555]	0.678* (0.472)	[-0.247, 1.603]	1.637* (7.105)	[-12.289, 15.563]
b17*LTEB	0.631* (0.141)	[0.355, 0.907]	1.833* (6.452)	[-10.813, 14.479]	-0.186 (18.639)	[-36.719, 36.346]	-0.170 (13.846)	[-27.308, 26.968]	0.734 (31.741)	[-61.478, 62.947]
b18*LTEB	1.231 (0.135)	[0.966, 1.496]	2.410* (1.120)	[0.215, 4.605]	0.420 (4.720)	[-8.831, 9.671]	0.433* (0.472)	[-0.492, 1.358]	1.312* (4.811)	[-8.117, 10.742]
b19*LTEB	1.291 (0.137)	[1.022, 1.560]	2.457* (0.135)	[2.193, 2.722]	0.470* (0.217)	[0.045, 0.895]	0.493* (0.472)	[-0.432, 1.418]	1.348* (0.573)	[0.225, 2.471]
b20*LTEB	0.969* (0.136)	[0.702, 1.236]	2.162* (1.341)	[-0.467, 4.790]	0.152* (0.896)	[-1.604, 1.909]	0.169* (0.472)	[-0.756, 1.094]	1.058* (1.444)	[-1.772, 3.888]
b21*LTEB	0.704* (0.135)	[0.439, 0.969]	1.909* (2.229)	[-2.460, 6.277]	-0.095 (19.267)	[-37.858, 37.669]	-0.098* (0.472)	[-1.023, 0.827]	0.817 (20.924)	[-40.194, 41.828]
b22*LTEB	0.688* (0.137)	[0.419, 0.957]	1.893* (1.341)	[-0.736, 4.521]	-0.126* (0.216)	[-0.549, 0.298]	-0.114* (0.472)	[-1.039, 0.811]	0.789* (1.599)	[-2.345, 3.923]
b23*LTEB	4.556* (0.136)	[4.289, 4.823]	5.768* (4.894)	[-3.824, 15.360]	3.750* (4.748)	[-5.556, 13.056]	3.748* (0.273)	[3.213, 4.283]	4.663* (6.729)	[-8.525, 17.852]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
b24*LTEB	1.489 (0.135)	[1.224, 1.754]	2.732* (4.673)	[-6.427, 11.891]	0.694 (9.662)	[-18.243, 19.632]	0.691* (0.273)	[0.156, 1.226]	1.661 (10.244)	[-18.418, 21.739]
b25*LTEB	1.661* (0.134)	[1.398, 1.924]	2.852* (2.007)	[-1.082, 6.786]	0.843* (2.111)	[-3.295, 4.981]	0.865* (0.472)	[-0.060, 1.790]	1.755* (2.560)	[-3.262, 6.773]
b26*LTEB	1.339 (0.139)	[1.067, 1.611]	2.518* (0.901)	[0.752, 4.284]	0.518* (0.218)	[0.091, 0.945]	0.541* (0.473)	[-0.386, 1.468]	1.412* (1.002)	[-0.552, 3.376]
b27*LTEB	1.228 (0.134)	[0.965, 1.491]	2.498* (6.452)	[-10.148, 15.144]	0.411* (2.145)	[-3.793, 4.615]	0.430* (0.370)	[-0.296, 1.155]	1.414 (8.284)	[-14.823, 17.650]
b28*LTEB	1.894* (0.135)	[1.629, 2.159]	3.115* (3.339)	[-3.429, 9.660]	1.110 (12.436)	[-23.265, 25.485]	1.029 (18.788)	[-35.796, 37.853]	1.884 (31.337)	[-59.536, 63.305]
b29*LTEB	1.616 (0.134)	[1.353, 1.879]	2.785* (0.900)	[1.021, 4.549]	0.799* (2.279)	[-3.668, 5.265]	0.819* (0.472)	[-0.106, 1.744]	1.683* (2.179)	[-2.587, 5.954]
b30*LTEB	0.916* (0.134)	[0.653, 1.179]	2.133* (3.562)	[-4.849, 9.114]	0.100* (2.279)	[-4.367, 4.566]	0.117* (0.471)	[-0.806, 1.040]	1.036* (4.333)	[-7.456, 9.529]
b31*LTEB	1.328 (0.135)	[1.063, 1.593]	2.536* (2.896)	[-3.140, 8.212]	0.507* (0.730)	[-0.924, 1.938]	0.532* (0.471)	[-0.392, 1.455]	1.440* (3.574)	[-5.565, 8.445]
b32*LTEB	4.891* (0.139)	[4.619, 5.163]	6.121* (4.229)	[-2.168, 14.410]	4.080 (5.756)	[-7.202, 15.362]	4.082* (0.472)	[3.157, 5.007]	5.015* (6.583)	[-7.888, 17.917]
b33*LTEB	0.517* (0.139)	[0.245, 0.789]	1.732* (1.342)	[-0.899, 4.362]	-0.294* (0.217)	[-0.719, 0.131]	-0.287* (0.473)	[-1.214, 0.640]	0.629* (1.599)	[-2.505, 3.763]
b34*LTEB	1.374 (0.134)	[1.111, 1.637]	2.556* (1.341)	[-0.073, 5.184]	0.553* (0.730)	[-0.878, 1.984]	0.577* (0.370)	[-0.149, 1.302]	1.450* (1.502)	[-1.494, 4.394]
b35*LTEB	0.985* (0.137)	[0.716, 1.254]	2.212* (4.894)	[-7.380, 11.804]	0.169* (1.932)	[-3.617, 3.956]	0.185* (0.472)	[-0.740, 1.110]	1.117* (6.094)	[-10.828, 13.061]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b36*LTEB	1.726* (0.136)	[1.459, 1.993]	2.968* (4.450)	[-5.754, 11.69]	0.935 (10.228)	[-19.112, 20.982]	0.874 (14.354)	[-27.26, 29.008]	1.765 (25.464)	[-48.144, 51.675]
b37*LTEB	1.714* (0.136)	[1.447, 1.981]	2.917* (2.673)	[-2.322, 8.156]	0.896* (2.481)	[-3.967, 5.759]	0.916* (0.273)	[0.381, 1.451]	1.823* (3.536)	[-5.107, 8.754]
b38*LTEB	1.293 (0.135)	[1.028, 1.558]	2.500* (2.896)	[-3.176, 8.176]	0.480* (3.847)	[-7.061, 8.020]	0.493* (0.135)	[0.229, 0.758]	1.404* (4.685)	[-7.778, 10.587]
b39*LTEB	0.924* (0.134)	[0.661, 1.187]	2.107* (1.120)	[-0.088, 4.302]	0.103* (0.133)	[-0.158, 0.363]	0.126* (0.471)	[-0.797, 1.049]	1.001* (1.390)	[-1.723, 3.725]
b40*LTEB	1.585 (0.135)	[1.320, 1.850]	2.790* (2.673)	[-2.449, 8.029]	0.801 (12.643)	[-23.979, 25.581]	0.787* (0.137)	[0.519, 1.056]	1.718 (13.457)	[-24.657, 28.094]
b41*LTEB	0.897* (0.134)	[0.634, 1.160]	2.098* (2.229)	[-2.271, 6.466]	0.080* (1.407)	[-2.678, 2.838]	0.097* (0.370)	[-0.629, 0.822]	0.995* (2.662)	[-4.222, 6.213]
b42*LTEB	1.812* (0.135)	[1.547, 2.077]	3.084* (5.785)	[-8.254, 14.423]	1.052 (19.642)	[-37.446, 39.550]	1.016* (0.472)	[0.091, 1.941]	2.044 (20.455)	[-38.047, 42.136]
b43*LTEB	1.052* (0.136)	[0.785, 1.319]	2.268* (3.562)	[-4.714, 9.249]	0.241 (4.356)	[-8.297, 8.778]	0.241 (4.572)	[-8.720, 9.203]	1.148 (10.273)	[-18.987, 21.283]
b44*LTEB	4.400* (0.136)	[4.133, 4.667]	5.537* (7.120)	[-8.418, 19.492]	3.586 (6.322)	[-8.805, 15.977]	3.593* (0.173)	[3.254, 3.932]	4.411 (9.941)	[-15.073, 23.896]
b45*LTEB	5.202* (0.139)	[4.930, 5.474]	6.442* (4.229)	[-1.847, 14.731]	4.393* (5.476)	[-6.340, 15.126]	4.394* (0.190)	[4.022, 4.767]	5.343* (6.643)	[-7.677, 18.363]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G22.

STEBvLTEB Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	0.280*** (0.059)	-	6.327*** (0.922)	-	7.124*** (1.330)	-	-2.941*** (0.489)	-	2.479 (2.148)	-
Intercept*LTEB	0.103 (0.115)	-	-0.168 (2.575)	-	0.231 (3.457)	-	0.766 (1.139)	-	-2.151 (5.524)	-
LEX	-	-	3.770*** (0.573)	-	-	-	-	-	1.220 (0.875)	-
LEX*LTEB	-	-	-0.183 (1.595)	-	-	-	-	-	0.139 (2.365)	-
NP	-	-	-	-	8.421*** (1.634)	-	-	-	2.749 (1.848)	-
NP*LTEB	-	-	-	-	0.138 (4.242)	-	-	-	-3.288 (5.119)	-
RC	-	-	-	-	-	-	-12.185*** (1.835)	-	-7.499** (2.532)	-
RC*LTEB	-	-	-	-	-	-	2.429 (4.315)	-	0.660 (5.809)	-
b01	-0.095 (0.079)	-	-14.548*** (2.201)	-	-9.389*** (1.806)	-	30.752*** (4.647)	-	11.187 (9.060)	-
b01*LTEB	-0.035 (0.153)	0.069	0.669 (6.127)	0.118	-0.188 (4.686)	0.095	-6.185 (10.927)	0.003	1.393 (21.649)	0.196
b02	-	-	-	-	-	-	-	-	-	-
b02*LTEB	-	-	-	-	-	-	-	-	-	-
b03	0.053 (0.079)	-	-6.770*** (1.041)	-	-8.888*** (1.737)	-	-0.922*** (0.167)	-	-5.661** (1.926)	-

b03*LTEB	-0.106 (0.154)	0.003	0.226 (2.896)	0.045	-0.253 (4.508)	0.023	0.088 (0.378)	0.070	3.189 (5.144)	0.269
b04	0.070 (0.079)	-	-7.557*** (1.164)	-	-3.407*** (0.679)	-	-0.904*** (0.167)	-	-4.117** (1.525)	-
b04*LTEB	-0.257+ (0.153)	0.157	0.111 (3.237)	0.113	-0.314 (1.759)	0.131	-0.064 (0.378)	0.226	0.868 (4.100)	0.430
b05	0.213** (0.080)	-	-6.609*** (1.041)	-	-1.935*** (0.424)	-	-0.761*** (0.167)	-	-3.279* (1.352)	-
b05*LTEB	-0.192 (0.154)	0.091	0.139 (2.896)	0.044	-0.226 (1.093)	0.063	0.002 (0.378)	0.158	0.446 (3.669)	0.361
b06	-0.362*** (0.078)	-	-3.166*** (0.435)	-	-9.792*** (1.832)	-	-1.338*** (0.167)	-	-4.939** (1.880)	-
b06*LTEB	-0.117 (0.152)	0.014	0.019 (1.201)	0.032	-0.272 (4.754)	0.012	0.078 (0.377)	0.081	3.515 (5.120)	0.283
b07	-0.551*** (0.078)	-	-3.754*** (0.495)	-	-3.185*** (0.517)	-	-1.528*** (0.167)	-	-3.039*** (0.643)	-
b07*LTEB	0.209 (0.152)	0.318	0.362 (1.369)	0.363	0.166 (1.337)	0.345	0.405 (0.377)	0.253	1.171 (1.728)	0.048
b08	-0.665*** (0.078)	-	-8.701*** (1.224)	-	-1.741*** (0.223)	-	-1.533*** (0.152)	-	-4.150* (1.683)	-
b08*LTEB	-0.168 (0.152)	0.066	0.225 (3.406)	0.016	-0.186 (0.564)	0.040	0.005 (0.342)	0.133	0.008 (4.589)	0.330
b09	0.161* (0.080)	-	-11.872*** (1.835)	-	-5.328*** (1.068)	-	9.784*** (1.452)	-	0.421 (4.311)	-
b09*LTEB	0.003 (0.155)	0.108	0.589 (5.107)	0.156	-0.087 (2.770)	0.135	-1.916 (3.412)	0.041	1.184 (10.646)	0.157
b10	0.774*** (0.084)	-	-11.225*** (1.835)	-	-0.803* (0.317)	-	10.503*** (1.469)	-	2.413 (4.329)	-
b10*LTEB	-0.205 (0.161)	0.104	0.362 (5.107)	0.076	-0.230 (0.809)	0.077	-2.144 (3.452)	0.169	-0.579 (10.676)	0.390
b11	0.141+ (0.080)	-	-1.068*** (0.199)	-	-3.285*** (0.670)	-	-0.833*** (0.167)	-	-1.969** (0.659)	-

b11*LTEB	0.020	0.126	0.079	0.173	-0.037	0.151	0.214	0.058	1.367	0.143
	(0.155)		(0.533)		(1.734)		(0.378)		(1.768)	
b12	0.426***	-	-11.618***	-	-11.35***	-	-0.544***	-	-7.895**	-
	(0.083)		(1.835)		(2.287)		(0.168)		(2.934)	
b12*LTEB	-0.637***	0.545	-0.072	0.519	-0.833	0.521	-0.446	0.615	3.540	0.845
	(0.156)		(5.107)		(5.936)		(0.378)		(7.782)	
b13	-0.005	-	-3.618***	-	-1.082***	-	-0.979***	-	-2.119**	-
	(0.079)		(0.555)		(0.223)		(0.167)		(0.664)	
b13*LTEB	-0.107	0.004	0.070	0.045	-0.125	0.022	0.088	0.070	0.235	0.271
	(0.153)		(1.539)		(0.564)		(0.378)		(1.835)	
b14	-0.333***	-	-7.562***	-	-6.630***	-	9.816***	-	1.525	-
	(0.078)		(1.103)		(1.225)		(1.530)		(3.566)	
b14*LTEB	-0.187	0.086	0.165	0.038	-0.290	0.059	-2.210	0.152	1.458	0.352
	(0.152)		(3.066)		(3.177)		(3.597)		(8.732)	
b15	0.061	-	-4.354***	-	-0.191*	-	-0.913***	-	-2.042*	-
	(0.079)		(0.676)		(0.093)		(0.167)		(0.893)	
b15*LTEB	-0.173	0.071	0.044	0.022	-0.177	0.045	0.021	0.139	-0.183	0.339
	(0.153)		(1.877)		(0.199)		(0.378)		(2.487)	
b16	0.158*	-	-8.268***	-	-8.950***	-	-0.816***	-	-6.120**	-
	(0.080)		(1.286)		(1.770)		(0.167)		(2.124)	
b16*LTEB	0.090	0.197	0.502	0.247	-0.060	0.223	0.285	0.131	3.394	0.067
	(0.156)		(3.576)		(4.593)		(0.379)		(5.651)	
b17	0.699***	-	-10.905***	-	-27.292***	-	30.541***	-	6.173	-
	(0.083)		(1.774)		(5.436)		(4.499)		(10.094)	
b17*LTEB	-0.123	0.020	0.430	0.014	-0.587	0.001	-6.070	0.081	8.767	0.287
	(0.161)		(4.936)		(14.111)		(10.578)		(25.651)	
b18	0.054	-	-1.954***	-	-7.034***	-	-0.920***	-	-3.506*	-
	(0.079)		(0.315)		(1.378)		(0.167)		(1.407)	
b18*LTEB	-0.095	0.008	0.005	0.058	-0.212	0.034	0.099	0.059	2.654	0.259
	(0.154)		(0.864)		(3.575)		(0.378)		(3.827)	
b19	-0.212**	-	-0.205**	-	-0.463***	-	-1.187***	-	-0.885***	-
	(0.079)		(0.077)		(0.092)		(0.167)		(0.218)	

b19*LTEB	0.465** (0.155)	0.580	0.451** (0.153)	0.613	0.459* (0.200)	0.604	0.661+ (0.379)	0.514	0.601 (0.483)	0.295
b20	0.045 (0.079)	-	-2.365*** (0.375)	-	-1.285*** (0.270)	-	-0.930*** (0.167)	-	-1.765*** (0.415)	-
b20*LTEB	0.055 (0.154)	0.161	0.172 (1.032)	0.209	0.033 (0.688)	0.188	0.250 (0.378)	0.095	0.538 (1.138)	0.108
b21	0.035 (0.079)	-	-3.981*** (0.616)	-	-28.909*** (5.619)	-	-0.939*** (0.167)	-	-11.305+ (6.045)	-
b21*LTEB	-0.037 (0.154)	0.067	0.159 (1.709)	0.116	-0.512 (14.586)	0.093	0.158 (0.378)	0.001	11.174 (16.627)	0.199
b22	0.413*** (0.081)	-	-2.006*** (0.375)	-	0.159+ (0.094)	-	-0.560*** (0.168)	-	-1.053* (0.442)	-
b22*LTEB	-0.224 (0.156)	0.123	-0.100 (1.032)	0.069	-0.228 (0.201)	0.097	-0.032 (0.379)	0.193	-0.156 (1.255)	0.387
b23	-0.028 (0.082)	-	-8.862*** (1.346)	-	-7.156*** (1.386)	-	-0.535*** (0.112)	-	-5.524** (2.023)	-
b23*LTEB	-0.116 (0.156)	0.013	0.315 (3.745)	0.036	-0.235 (3.596)	0.011	-0.018 (0.238)	0.084	2.377 (5.351)	0.275
b24	-0.159* (0.079)	-	-8.592*** (1.286)	-	-14.674*** (2.818)	-	-0.671*** (0.110)	-	-7.931** (3.040)	-
b24*LTEB	0.036 (0.153)	0.142	0.449 (3.576)	0.192	-0.202 (7.316)	0.168	0.139 (0.237)	0.076	5.427 (8.160)	0.121
b25	0.279*** (0.080)	-	-3.339*** (0.556)	-	-2.887*** (0.619)	-	-0.694*** (0.167)	-	-2.520*** (0.757)	-
b25*LTEB	-0.477** (0.153)	0.382	-0.295 (1.539)	0.328	-0.528 (1.602)	0.354	-0.284 (0.378)	0.450	0.684 (2.029)	0.645
b26	0.158* (0.080)	-	-1.454*** (0.257)	-	-0.095 (0.093)	-	-0.817*** (0.167)	-	-1.046*** (0.277)	-
b26*LTEB	0.254 (0.157)	0.364	0.328 (0.698)	0.408	0.249 (0.202)	0.390	0.449 (0.379)	0.298	0.341 (0.788)	0.090
b27	-0.460*** (0.078)	-	-12.115*** (1.773)	-	-3.674*** (0.629)	-	-1.205*** (0.137)	-	-5.736* (2.439)	-

b27*LTEB	-0.070 (0.152)	0.034	0.499 (4.936)	0.084	-0.123 (1.628)	0.060	0.079 (0.304)	0.033	0.802 (6.557)	0.229
b28	0.356*** (0.081)	-	-5.664*** (0.920)	-	-18.322*** (3.627)	-	40.862*** (6.104)	-	17.216+ (10.416)	-
b28*LTEB	-0.519*** (0.154)	0.425	-0.226 (2.557)	0.377	-0.829 (9.415)	0.402	-8.581 (14.352)	0.474	4.391 (25.502)	0.661
b29	-0.543*** (0.078)	-	-2.139*** (0.257)	-	-3.960*** (0.668)	-	-1.520*** (0.167)	-	-2.765*** (0.648)	-
b29*LTEB	0.284+ (0.152)	0.395	0.355 (0.698)	0.435	0.227 (1.729)	0.421	0.480 (0.378)	0.330	1.606 (1.735)	0.120
b30	-0.379*** (0.078)	-	-6.806*** (0.981)	-	-3.796*** (0.668)	-	-1.356*** (0.167)	-	-4.169** (1.276)	-
b30*LTEB	0.111 (0.152)	0.218	0.426 (2.727)	0.269	0.054 (1.729)	0.244	0.307 (0.377)	0.153	1.268 (3.428)	0.044
b31	-0.201* (0.079)	-	-5.421*** (0.798)	-	-1.278*** (0.223)	-	-1.177*** (0.167)	-	-2.834** (1.025)	-
b31*LTEB	0.040 (0.153)	0.146	0.294 (2.218)	0.194	0.022 (0.564)	0.172	0.236 (0.378)	0.081	0.324 (2.817)	0.120
b32	-0.084 (0.082)	-	-7.718*** (1.163)	-	-8.727*** (1.679)	-	-1.055*** (0.167)	-	-5.980** (1.966)	-
b32*LTEB	0.285+ (0.158)	0.396	0.662 (3.236)	0.450	0.141 (4.359)	0.420	0.478 (0.379)	0.328	3.452 (5.236)	0.147
b33	0.742*** (0.083)	-	-1.684*** (0.376)	-	0.487*** (0.096)	-	-0.229 (0.169)	-	-0.731+ (0.443)	-
b33*LTEB	-0.367* (0.159)	0.269	-0.238 (1.033)	0.210	-0.369+ (0.203)	0.241	-0.174 (0.38)	0.338	-0.293 (1.255)	0.526
b34	-0.117 (0.079)	-	-2.523*** (0.375)	-	-1.193*** (0.223)	-	-0.860*** (0.137)	-	-1.698*** (0.433)	-
b34*LTEB	-0.296+ (0.152)	0.197	-0.175 (1.032)	0.146	-0.314 (0.564)	0.171	-0.148 (0.304)	0.264	0.080 (1.185)	0.462
b35	0.199* (0.080)	-	-8.624*** (1.346)	-	-2.697*** (0.568)	-	-0.774*** (0.167)	-	-4.177* (1.785)	-

b35*LTEB	-0.012 (0.155)	0.093	0.417 (3.745)	0.140	-0.060 (1.467)	0.119	0.183 (0.379)	0.027	0.848 (4.818)	0.174
b36	0.324*** (0.080)	-	-7.697*** (1.224)	-	-15.038*** (2.983)	-	31.270*** (4.663)	-	11.748 (8.486)	-
b36*LTEB	-0.220 (0.155)	0.119	0.167 (3.406)	0.075	-0.474 (7.744)	0.095	-6.386 (10.966)	0.180	3.817 (20.728)	0.378
b37	-0.263*** (0.078)	-	-5.080*** (0.737)	-	-3.983*** (0.726)	-	-0.776*** (0.110)	-	-3.344** (1.058)	-
b37*LTEB	0.255+ (0.154)	0.365	0.488 (2.047)	0.412	0.193 (1.881)	0.391	0.359 (0.238)	0.300	1.560 (2.810)	0.099
b38	0.070 (0.079)	-	-5.151*** (0.798)	-	-5.705*** (1.124)	-	0.070 (0.080)	-	-3.495* (1.413)	-
b38*LTEB	-0.113 (0.154)	0.010	0.141 (2.218)	0.037	-0.208 (2.914)	0.016	-0.113 (0.154)	0.077	1.951 (3.736)	0.278
b39	-0.587*** (0.078)	-	-2.582*** (0.315)	-	-0.584*** (0.078)	-	-1.564*** (0.167)	-	-1.821*** (0.381)	-
b39*LTEB	0.068 (0.152)	0.175	0.162 (0.864)	0.218	0.067 (0.152)	0.200	0.264 (0.377)	0.109	0.045 (1.091)	0.096
b40	-0.090 (0.079)	-	-4.907*** (0.737)	-	-19.083*** (3.687)	-	-0.139+ (0.079)	-	-7.871* (3.912)	-
b40*LTEB	-0.019 (0.153)	0.086	0.216 (2.047)	0.134	-0.330 (9.572)	0.113	-0.008 (0.154)	0.020	7.227 (10.704)	0.179
b41	-0.060 (0.079)	-	-4.075*** (0.616)	-	-2.164*** (0.416)	-	-0.804*** (0.137)	-	-2.497*** (0.781)	-
b41*LTEB	-0.147 (0.153)	0.045	0.049 (1.709)	0.003	-0.182 (1.071)	0.019	0.001 (0.304)	0.112	0.569 (2.105)	0.312
b42	0.047 (0.079)	-	-10.385*** (1.590)	-	-29.46*** (5.728)	-	-0.927*** (0.167)	-	-13.536* (5.985)	-
b42*LTEB	-0.157 (0.154)	0.055	0.347 (4.426)	0.011	-0.643 (14.870)	0.031	0.037 (0.378)	0.122	11.032 (16.278)	0.327
b43	-0.245** (0.079)	-	-6.670*** (0.981)	-	-6.785*** (1.272)	-	9.612*** (1.486)	-	1.614 (3.396)	-

b43*LTEB	0.265+ (0.154)	0.376	0.579 (2.727)	0.425	0.158 (3.300)	0.402	-1.701 (3.494)	0.308	2.048 (8.344)	0.106
b44	0.072 (0.082)	-	-12.784*** (1.957)	-	-9.422*** (1.844)	-	0.309*** (0.089)	-	-7.037* (3.012)	-
b44*LTEB	-0.488** (0.155)	0.393	0.127 (5.447)	0.355	-0.646 (4.787)	0.369	-0.537** (0.176)	0.463	2.723 (7.919)	0.673
b45	-1.841*** (0.085)	-	-9.483*** (1.164)	-	-10.063*** (1.598)	-	-2.120*** (0.095)	-	-7.190*** (1.998)	-
b45*LTEB	0.563*** (0.161)	0.680	0.935 (3.236)	0.728	0.426 (4.147)	0.704	0.617** (0.188)	0.611	3.514 (5.292)	0.416
delta1	4.183*** (0.033)	-	4.201*** (0.033)	-	4.182*** (0.033)	-	4.189*** (0.033)	-	4.225*** (0.033)	-
delta1*LTEB	-0.017 (0.066)	-	-0.023 (0.066)	-	-0.016 (0.066)	-	-0.022 (0.066)	-	-0.034 (0.066)	-
delta2	5.900*** (0.072)	-	5.928*** (0.073)	-	5.900*** (0.072)	-	5.907*** (0.072)	-	5.963*** (0.073)	-
delta2*LTEB	0.490** (0.184)	-	0.479** (0.184)	-	0.491** (0.184)	-	0.485** (0.184)	-	0.463* (0.184)	-
delta3	7.082*** (0.129)	-	7.114*** (0.130)	-	7.082*** (0.129)	-	7.090*** (0.129)	-	7.152*** (0.130)	-
delta3*LTEB	0.500 (0.331)	-	0.487 (0.332)	-	0.500 (0.331)	-	0.494 (0.331)	-	0.467 (0.333)	-
Intercept Variance	0.497		0.497		0.489		0.482		0.501	
LEX Variance	-		0.018		-		-		0.038	
NP Variance	-		-		0.001		-		0.002	
RC Variance	-		-		-		0.004		0.021	
Intercept*Feature Covariance	-		0.093		0.019		-0.039		See Table G30	

Note: + p < .10, * p < .05, ** p < .01, *** p < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G23.

STEBvLTEB Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b01*LTEB	0.068 (0.153)	[-0.232, 0.368]	0.116 (6.127)	[-11.893, 12.125]	0.093 (4.686)	[-9.091, 9.278]	0.003 (10.927)	[-21.414, 21.419]	-0.192 (21.649)	[-42.624, 42.240]
b02*LTEB	-	-	-	-	-	-	-	-	-	-
b03*LTEB	-0.003 (0.154)	[-0.305, 0.299]	0.044 (2.896)	[-5.632, 5.720]	0.023 (4.508)	[-8.813, 8.858]	-0.069* (0.378)	[-0.810, 0.672]	-0.264 (5.144)	[-10.346, 9.818]
b04*LTEB	-0.154 (0.153)	[-0.454, 0.146]	-0.110 (3.237)	[-6.455, 6.234]	-0.128 (1.759)	[-3.576, 3.320]	-0.221* (0.378)	[-0.962, 0.520]	-0.421 (4.100)	[-8.457, 7.615]
b05*LTEB	-0.089 (0.154)	[-0.391, 0.213]	-0.043 (2.896)	[-5.719, 5.633]	-0.062 (1.093)	[-2.204, 2.080]	-0.155* (0.378)	[-0.896, 0.586]	-0.354 (3.669)	[-7.545, 6.838]
b06*LTEB	-0.014 (0.152)	[-0.312, 0.284]	0.032 (1.201)	[-2.322, 2.386]	0.012 (4.754)	[-9.306, 9.329]	-0.079* (0.377)	[-0.818, 0.660]	-0.277 (5.120)	[-10.312, 9.758]
b07*LTEB	0.312 (0.152)	[0.014, 0.610]	0.355 (1.369)	[-2.328, 3.039]	0.338 (1.337)	[-2.282, 2.959]	0.248 (0.377)	[-0.491, 0.987]	0.047 (1.728)	[-3.340, 3.434]
b08*LTEB	-0.065 (0.152)	[-0.363, 0.233]	-0.016 (3.406)	[-6.692, 6.660]	-0.039 (0.564)	[-1.145, 1.066]	-0.130* (0.342)	[-0.800, 0.540]	-0.324 (4.589)	[-9.318, 8.671]
b09*LTEB	0.106 (0.155)	[-0.198, 0.410]	0.153 (5.107)	[-9.857, 10.163]	0.132 (2.770)	[-5.297, 5.561]	0.040 (3.412)	[-6.647, 6.728]	-0.154 (10.646)	[-21.020, 20.712]
b10*LTEB	-0.102 (0.161)	[-0.418, 0.214]	-0.074 (5.107)	[-10.084, 9.936]	-0.075 (0.809)	[-1.661, 1.510]	-0.166 (3.452)	[-6.932, 6.600]	-0.382 (10.676)	[-21.307, 20.543]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b11*LTEB	0.123 (0.155)	[-0.181, 0.427]	0.170 (0.533)	[-0.875, 1.215]	0.148 (1.734)	[-3.250, 3.547]	0.057 (0.378)	[-0.684, 0.798]	-0.140 (1.768)	[-3.605, 3.325]
b12*LTEB	-0.534* (0.156)	[-0.840, -0.228]	-0.508 (5.107)	[-10.518, 9.502]	-0.511 (5.936)	[-12.145, 11.124]	-0.603* (0.378)	[-1.344, 0.138]	-0.828 (7.782)	[-16.081, 14.424]
b13*LTEB	-0.004 (0.153)	[-0.304, 0.296]	0.044 (1.539)	[-2.973, 3.060]	0.022 (0.564)	[-1.084, 1.127]	-0.069* (0.378)	[-0.810, 0.672]	-0.266 (1.835)	[-3.862, 3.331]
b14*LTEB	-0.084 (0.152)	[-0.382, 0.214]	-0.037 (3.066)	[-6.046, 5.973]	-0.058 (3.177)	[-6.285, 6.169]	-0.149 (3.597)	[-7.199, 6.901]	-0.345 (8.732)	[-17.460, 16.770]
b15*LTEB	-0.070 (0.153)	[-0.370, 0.230]	-0.021 (1.877)	[-3.700, 3.658]	-0.044 (0.199)	[-0.434, 0.346]	-0.136* (0.378)	[-0.877, 0.605]	-0.332 (2.487)	[-5.206, 4.543]
b16*LTEB	0.193 (0.156)	[-0.113, 0.499]	0.242 (3.576)	[-6.767, 7.251]	0.218 (4.593)	[-8.784, 9.221]	0.128 (0.379)	[-0.615, 0.871]	-0.065 (5.651)	[-11.141, 11.011]
b17*LTEB	-0.020 (0.161)	[-0.336, 0.296]	0.013 (4.936)	[-9.661, 9.688]	0.001 (14.111)	[-27.657, 27.659]	-0.079 (10.578)	[-20.812, 20.654]	-0.282 (25.651)	[-50.558, 49.994]
b18*LTEB	0.008 (0.154)	[-0.294, 0.310]	0.057 (0.864)	[-1.636, 1.750]	0.033 (3.575)	[-6.974, 7.040]	-0.058* (0.378)	[-0.799, 0.683]	-0.254 (3.827)	[-7.754, 7.247]
b19*LTEB	0.568* (0.155)	[0.264, 0.872]	0.601* (0.153)	[0.301, 0.900]	0.592 (0.200)	[0.200, 0.984]	0.504 (0.379)	[-0.239, 1.247]	0.289* (0.483)	[-0.657, 1.236]
b20*LTEB	0.158 (0.154)	[-0.144, 0.460]	0.204 (1.032)	[-1.818, 2.227]	0.184 (0.688)	[-1.165, 1.532]	0.093 (0.378)	[-0.648, 0.834]	-0.106 (1.138)	[-2.336, 2.125]
b21*LTEB	0.066 (0.154)	[-0.236, 0.368]	0.113 (1.709)	[-3.236, 3.463]	0.091 (14.586)	[-28.497, 28.680]	0.001* (0.378)	[-0.740, 0.742]	-0.195 (16.627)	[-32.784, 32.394]
b22*LTEB	-0.121 (0.156)	[-0.427, 0.185]	-0.068 (1.032)	[-2.090, 1.955]	-0.095 (0.201)	[-0.489, 0.299]	-0.189* (0.379)	[-0.932, 0.554]	-0.379 (1.255)	[-2.839, 2.081]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b23*LTEB	-0.013 (0.156)	[-0.319, 0.293]	0.035 (3.745)	[-7.305, 7.375]	0.011 (3.596)	[-7.037, 7.059]	-0.083* (0.238)	[-0.549, 0.384]	-0.270 (5.351)	[-10.758, 10.218]
b24*LTEB	0.139 (0.153)	[-0.161, 0.439]	0.189 (3.576)	[-6.820, 7.198]	0.165 (7.316)	[-14.174, 14.504]	0.074* (0.237)	[-0.390, 0.539]	-0.118 (8.160)	[-16.112, 15.875]
b25*LTEB	-0.374* (0.153)	[-0.674, -0.074]	-0.321 (1.539)	[-3.338, 2.695]	-0.347 (1.602)	[-3.487, 2.793]	-0.441* (0.378)	[-1.182, 0.300]	-0.632 (2.029)	[-4.609, 3.345]
b26*LTEB	0.357 (0.157)	[0.049, 0.665]	0.399 (0.698)	[-0.969, 1.767]	0.382 (0.202)	[-0.014, 0.778]	0.292 (0.379)	[-0.451, 1.035]	0.089* (0.788)	[-1.456, 1.633]
b27*LTEB	0.033 (0.152)	[-0.265, 0.331]	0.082 (4.936)	[-9.592, 9.757]	0.059 (1.628)	[-3.132, 3.250]	-0.032* (0.304)	[-0.628, 0.564]	-0.225 (6.557)	[-13.076, 12.627]
b28*LTEB	-0.416* (0.154)	[-0.718, -0.114]	-0.369 (2.557)	[-5.381, 4.642]	-0.394 (9.415)	[-18.847, 18.060]	-0.465 (14.352)	[-28.595, 27.665]	-0.648 (25.502)	[-50.632, 49.336]
b29*LTEB	0.387 (0.152)	[0.089, 0.685]	0.426 (0.698)	[-0.942, 1.794]	0.412 (1.729)	[-2.977, 3.801]	0.323 (0.378)	[-0.418, 1.064]	0.117 (1.735)	[-3.283, 3.518]
b30*LTEB	0.214 (0.152)	[-0.084, 0.512]	0.263 (2.727)	[-5.082, 5.608]	0.239 (1.729)	[-3.150, 3.628]	0.150 (0.377)	[-0.589, 0.889]	-0.043 (3.428)	[-6.762, 6.676]
b31*LTEB	0.143 (0.153)	[-0.157, 0.443]	0.190 (2.218)	[-4.158, 4.537]	0.169 (0.564)	[-0.937, 1.274]	0.079 (0.378)	[-0.662, 0.820]	-0.117 (2.817)	[-5.639, 5.404]
b32*LTEB	0.388 (0.158)	[0.078, 0.698]	0.441 (3.236)	[-5.902, 6.783]	0.412 (4.359)	[-8.132, 8.955]	0.321 (0.379)	[-0.422, 1.064]	0.144 (5.236)	[-10.119, 10.406]
b33*LTEB	-0.264* (0.159)	[-0.576, 0.048]	-0.206 (1.033)	[-2.230, 1.819]	-0.236* (0.203)	[-0.634, 0.162]	-0.331* (0.380)	[-1.076, 0.414]	-0.516 (1.255)	[-2.976, 1.944]
b34*LTEB	-0.193 (0.152)	[-0.491, 0.105]	-0.143 (1.032)	[-2.165, 1.880]	-0.167 (0.564)	[-1.273, 0.938]	-0.259* (0.304)	[-0.855, 0.337]	-0.452 (1.185)	[-2.775, 1.870]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b35*LTEB	0.091 (0.155)	[-0.213, 0.395]	0.137 (3.745)	[-7.203, 7.477]	0.116 (1.467)	[-2.759, 2.992]	0.026 (0.379)	[-0.717, 0.769]	-0.170 (4.818)	[-9.613, 9.273]
b36*LTEB	-0.117 (0.155)	[-0.421, 0.187]	-0.074 (3.406)	[-6.750, 6.602]	-0.093 (7.744)	[-15.271, 15.085]	-0.177 (10.966)	[-21.670, 21.317]	-0.370 (20.728)	[-40.997, 40.256]
b37*LTEB	0.358 (0.154)	[0.056, 0.660]	0.403 (2.047)	[-3.609, 4.415]	0.383 (1.881)	[-3.304, 4.070]	0.294* (0.238)	[-0.172, 0.761]	0.097 (2.810)	[-5.411, 5.604]
b38*LTEB	-0.010 (0.154)	[-0.312, 0.292]	0.037 (2.218)	[-4.311, 4.384]	0.016 (2.914)	[-5.696, 5.727]	-0.076* (0.154)	[-0.378, 0.226]	-0.272 (3.736)	[-7.595, 7.050]
b39*LTEB	0.171 (0.152)	[-0.127, 0.469]	0.214 (0.864)	[-1.479, 1.907]	0.196 (0.152)	[-0.102, 0.494]	0.107 (0.377)	[-0.632, 0.846]	-0.094 (1.091)	[-2.232, 2.044]
b40*LTEB	0.084 (0.153)	[-0.216, 0.384]	0.131 (2.047)	[-3.881, 4.143]	0.110 (9.572)	[-18.651, 18.871]	0.020* (0.154)	[-0.282, 0.321]	-0.176 (10.704)	[-21.156, 20.804]
b41*LTEB	-0.044 (0.153)	[-0.344, 0.256]	0.003 (1.709)	[-3.346, 3.353]	-0.018 (1.071)	[-2.118, 2.081]	-0.110* (0.304)	[-0.706, 0.486]	-0.305 (2.105)	[-4.431, 3.821]
b42*LTEB	-0.054 (0.154)	[-0.356, 0.248]	-0.011 (4.426)	[-8.686, 8.664]	-0.030 (14.870)	[-29.175, 29.115]	-0.120* (0.378)	[-0.861, 0.621]	-0.320 (16.278)	[-32.225, 31.585]
b43*LTEB	0.368 (0.154)	[0.066, 0.670]	0.416 (2.727)	[-4.929, 5.761]	0.394 (3.300)	[-6.074, 6.862]	0.301 (3.494)	[-6.547, 7.150]	0.104 (8.344)	[-16.250, 16.458]
b44*LTEB	-0.385* (0.155)	[-0.689, -0.081]	-0.348 (5.447)	[-11.024, 10.328]	-0.361 (4.787)	[-9.744, 9.021]	-0.454* (0.176)	[-0.799, -0.109]	-0.659 (7.919)	[-16.180, 14.862]
b45*LTEB	0.666* (0.161)	[0.350, 0.982]	0.714 (3.236)	[-5.629, 7.056]	0.690 (4.147)	[-7.438, 8.818]	0.598 (0.188)	[0.230, 0.967]	0.408 (5.292)	[-9.964, 10.780]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G24.

EPvSPA Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	-0.997*** (0.020)	-	21.182*** (1.031)	-	15.827*** (1.219)	-	-8.505*** (0.432)	-	13.485*** (1.652)	-
Intercept*SPA	1.424*** (0.071)	-	-15.784*** (1.756)	-	-8.522*** (2.413)	-	6.724*** (0.850)	-	-10.951** (3.589)	-
LEX	-	-	13.686*** (0.637)	-	-	-	-	-	8.551*** (0.793)	-
LEX*SPA	-	-	-10.586*** (1.089)	-	-	-	-	-	-7.734*** (1.675)	-
NP	-	-	-	-	20.606*** (1.494)	-	-	-	5.230*** (1.513)	-
NP*SPA	-	-	-	-	-12.141*** (2.963)	-	-	-	-2.655 (3.417)	-
RC	-	-	-	-	-	-	-28.799*** (1.657)	-	-14.071*** (1.754)	-
RC*SPA	-	-	-	-	-	-	20.441*** (3.218)	-	9.217* (4.115)	-
b01	-0.369*** (0.027)	-	-53.051*** (2.447)	-	-23.133*** (1.650)	-	72.593*** (4.196)	-	-3.418 (6.425)	-
b01*SPA	0.262** (0.090)	1.721	41.070*** (4.183)	3.064	13.685*** (3.273)	0.747	-51.545*** (8.149)	0.820	9.642 (14.994)	2.056
b02	-	-	-	-	-	-	-	-	-	-
b02*SPA	-	-	-	-	-	-	-	-	-	-
b03	0.344*** (0.026)	-	-24.517*** (1.155)	-	-21.551*** (1.587)	-	-1.959*** (0.135)	-	-21.904*** (1.614)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b03*SPA	-0.382*** (0.090)	1.063	18.873*** (1.977)	2.310	12.527*** (3.148)	0.085	1.253*** (0.273)	0.214	17.285*** (3.287)	1.399
b04	0.079** (0.026)	-	-27.720*** (1.292)	-	-8.436*** (0.618)	-	-2.225*** (0.135)	-	-20.615*** (1.402)	-
b04*SPA	-0.154+ (0.090)	1.296	21.378*** (2.210)	2.554	4.868*** (1.227)	0.310	1.483*** (0.273)	0.448	17.456*** (2.781)	1.643
b05	0.478*** (0.026)	-	-24.381*** (1.155)	-	-4.783*** (0.382)	-	-1.824*** (0.135)	-	-17.542*** (1.263)	-
b05*SPA	-0.364*** (0.091)	1.082	18.890*** (1.977)	2.327	2.739*** (0.761)	0.095	1.270*** (0.273)	0.231	15.152*** (2.526)	1.409
b06	-0.076** (0.027)	-	-10.32*** (0.477)	-	-23.169*** (1.674)	-	-2.382*** (0.135)	-	-13.486*** (1.528)	-
b06*SPA	-0.323*** (0.090)	1.124	7.617*** (0.818)	2.339	13.292*** (3.320)	0.147	1.313*** (0.273)	0.275	9.220** (3.347)	1.425
b07	0.060* (0.026)	-	-11.639*** (0.544)	-	-6.394*** (0.468)	-	-2.245*** (0.135)	-	-10.032*** (0.591)	-
b07*SPA	-0.606*** (0.089)	0.835	8.462*** (0.933)	2.056	3.201*** (0.932)	0.152	1.029*** (0.273)	0.015	7.611*** (1.125)	1.133
b08	-0.952*** (0.028)	-	-30.219*** (1.359)	-	-3.591*** (0.193)	-	-2.999*** (0.121)	-	-20.957*** (1.565)	-
b08*SPA	0.295** (0.090)	1.754	22.959*** (2.325)	3.023	1.854*** (0.390)	0.766	1.746*** (0.246)	0.905	17.895*** (3.211)	2.111
b09	-0.005 (0.026)	-	-43.893*** (2.039)	-	-13.448*** (0.974)	-	22.757*** (1.309)	-	-19.754*** (3.284)	-
b09*SPA	0.096 (0.091)	1.551	34.094*** (3.487)	2.859	8.022*** (1.934)	0.568	-16.068*** (2.544)	0.686	19.419** (7.377)	1.924
b10	0.910*** (0.026)	-	-42.916*** (2.039)	-	-2.954*** (0.281)	-	23.923*** (1.324)	-	-16.227*** (3.360)	-
b10*SPA	-0.174+ (0.096)	1.276	33.789*** (3.487)	2.548	2.103*** (0.562)	0.289	-16.517*** (2.573)	0.415	17.797* (7.608)	1.613

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b11	-0.296*** (0.027)	-	-4.678*** (0.206)	-	-8.687*** (0.609)	-	-2.602*** (0.135)	-	-6.298*** (0.544)	-
b11*SPA	0.535*** (0.092)	1.999	3.921*** (0.360)	3.180	5.481*** (1.209)	1.010	2.173*** (0.273)	1.153	4.837*** (1.175)	2.254
b12	-3.938*** (0.026)	-	-47.728*** (2.039)	-	-32.767*** (2.090)	-	-6.232*** (0.135)	-	-39.718*** (2.505)	-
b12*SPA	4.336*** (0.095)	5.879	38.234*** (3.487)	7.084	21.328*** (4.146)	4.892	5.963*** (0.274)	5.021	33.557*** (5.028)	6.145
b13	-0.303*** (0.027)	-	-13.468*** (0.612)	-	-2.941*** (0.193)	-	-2.609*** (0.135)	-	-10.344*** (0.645)	-
b13*SPA	0.363*** (0.091)	1.824	10.558*** (1.049)	3.039	1.917*** (0.390)	0.830	2.001*** (0.273)	0.977	8.911*** (1.259)	2.116
b14	-0.268*** (0.027)	-	-26.605*** (1.224)	-	-15.690*** (1.118)	-	23.735*** (1.381)	-	-8.924*** (2.576)	-
b14*SPA	-0.031 (0.090)	1.422	20.368*** (2.093)	2.680	9.063*** (2.218)	0.441	-17.074*** (2.682)	0.556	9.189 (5.902)	1.731
b15	0.012 (0.026)	-	-16.076*** (0.748)	-	-0.606*** (0.052)	-	-2.292*** (0.135)	-	-11.345*** (0.854)	-
b15*SPA	0.021 (0.091)	1.475	12.481*** (1.281)	2.701	0.385** (0.127)	0.481	1.657*** (0.273)	0.626	9.967*** (1.738)	1.778
b16	0.163*** (0.026)	-	-30.555*** (1.427)	-	-22.145*** (1.617)	-	-2.141*** (0.135)	-	-25.860*** (1.808)	-
b16*SPA	0.037 (0.092)	1.491	23.833*** (2.441)	2.759	13.190*** (3.208)	0.514	1.674*** (0.273)	0.643	21.093*** (3.631)	1.861
b17	1.337*** (0.026)	-	-40.994*** (1.971)	-	-67.181*** (4.969)	-	71.905*** (4.062)	-	-8.060 (7.239)	-
b17*SPA	-0.786*** (0.094)	0.651	32.003*** (3.370)	1.881	39.602*** (9.856)	0.336	-50.889*** (7.888)	0.200	9.440 (16.596)	0.963
b18	0.114*** (0.026)	-	-7.199*** (0.341)	-	-17.246*** (1.258)	-	-2.191*** (0.135)	-	-10.005*** (1.147)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b18*SPA	-0.063 (0.091)	1.389	5.601*** (0.587)	2.594	10.173*** (2.497)	0.409	1.574*** (0.273)	0.541	7.071** (2.515)	1.674
b19	0.351*** (0.026)	-	0.334*** (0.025)	-	-0.271*** (0.052)	-	-1.952*** (0.135)	-	-0.953*** (0.147)	-
b19*SPA	-0.469*** (0.090)	0.975	-0.449*** (0.089)	2.178	-0.100 (0.126)	0.014	1.165*** (0.273)	0.124	0.374 (0.351)	1.247
b20	0.517*** (0.026)	-	-8.275*** (0.408)	-	-2.744*** (0.237)	-	-1.784*** (0.135)	-	-6.948*** (0.409)	-
b20*SPA	-0.530*** (0.090)	0.912	6.283*** (0.703)	2.134	1.395** (0.477)	0.075	1.103*** (0.273)	0.061	5.625*** (0.757)	1.207
b21	0.680*** (0.026)	-	-13.964*** (0.680)	-	-70.177*** (5.137)	-	-1.621*** (0.135)	-	-27.601*** (4.908)	-
b21*SPA	-0.676*** (0.091)	0.763	10.668*** (1.166)	1.996	41.091*** (10.188)	0.204	0.956*** (0.273)	0.089	17.516+ (10.975)	1.085
b22	0.887*** (0.026)	-	-7.918*** (0.408)	-	0.259*** (0.052)	-	-1.412*** (0.135)	-	-5.922*** (0.440)	-
b22*SPA	-0.535*** (0.093)	0.907	6.284*** (0.703)	2.135	-0.163 (0.128)	0.079	1.097*** (0.274)	0.055	5.286*** (0.865)	1.208
b23	-3.317*** (0.027)	-	-35.467*** (1.495)	-	-20.783*** (1.266)	-	-4.517*** (0.075)	-	-28.446*** (1.750)	-
b23*SPA	3.561*** (0.095)	5.088	28.458*** (2.557)	6.334	13.862*** (2.511)	4.112	4.413*** (0.165)	4.232	24.419*** (3.503)	5.413
b24	-0.228*** (0.027)	-	-30.955*** (1.427)	-	-35.776*** (2.576)	-	-1.439*** (0.075)	-	-29.095*** (2.473)	-
b24*SPA	0.103 (0.090)	1.558	23.903*** (2.441)	2.830	21.063*** (5.110)	0.594	0.963*** (0.163)	0.711	22.525*** (5.206)	1.940
b25	-0.475*** (0.027)	-	-13.639*** (0.612)	-	-8.226*** (0.562)	-	-2.781*** (0.135)	-	-11.816*** (0.684)	-
b25*SPA	0.725*** (0.092)	2.193	10.918*** (1.050)	3.407	5.294*** (1.118)	1.203	2.364*** (0.273)	1.348	9.935*** (1.316)	2.489

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b26	0.463*** (0.026)	-	-5.409*** (0.273)	-	-0.160** (0.052)	-	-1.839*** (0.135)	-	-4.506*** (0.282)	-
b26*SPA	-0.201* (0.092)	1.248	4.345*** (0.474)	2.457	0.167 (0.128)	0.258	1.434*** (0.273)	0.398	3.952*** (0.535)	1.528
b27	-0.377*** (0.027)	-	-42.813*** (1.971)	-	-8.252*** (0.571)	-	-2.136*** (0.105)	-	-29.815*** (2.229)	-
b27*SPA	-0.078 (0.090)	1.374	32.782*** (3.370)	2.676	4.567*** (1.136)	0.387	1.170*** (0.216)	0.525	25.571*** (4.543)	1.774
b28	-0.672*** (0.027)	-	-22.623*** (1.020)	-	-46.430*** (3.315)	-	95.172*** (5.511)	-	20.895** (7.264)	-
b28*SPA	0.931*** (0.092)	2.403	17.935*** (1.745)	3.654	27.916*** (6.576)	1.455	-67.135*** (10.704)	1.473	-11.533 (16.927)	2.573
b29	-0.492*** (0.027)	-	-6.341*** (0.273)	-	-8.862*** (0.607)	-	-2.798*** (0.135)	-	-7.407*** (0.540)	-
b29*SPA	0.091 (0.090)	1.546	4.628*** (0.474)	2.746	5.027*** (1.206)	0.559	1.728*** (0.273)	0.699	5.237*** (1.137)	1.820
b30	0.198*** (0.026)	-	-23.212*** (1.088)	-	-8.173*** (0.607)	-	-2.106*** (0.135)	-	-17.71*** (1.173)	-
b30*SPA	-0.504*** (0.090)	0.939	17.627*** (1.861)	2.183	4.433*** (1.206)	0.047	1.131*** (0.273)	0.089	14.593*** (2.304)	1.266
b31	-0.105*** (0.027)	-	-19.129*** (0.884)	-	-2.744*** (0.193)	-	-2.411*** (0.135)	-	-13.815*** (0.974)	-
b31*SPA	0.006 (0.090)	1.459	14.741*** (1.513)	2.695	1.562*** (0.390)	0.468	1.643*** (0.273)	0.612	11.881*** (1.954)	1.777
b32	-3.304*** (0.027)	-	-31.105*** (1.292)	-	-24.475*** (1.535)	-	-5.596*** (0.135)	-	-27.195*** (1.668)	-
b32*SPA	3.570*** (0.095)	5.097	25.103*** (2.210)	6.356	16.053*** (3.044)	4.117	5.196*** (0.274)	4.238	22.795*** (3.358)	5.428
b33	1.249*** (0.026)	-	-7.572*** (0.408)	-	0.615*** (0.052)	-	-1.048*** (0.135)	-	-5.576*** (0.440)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b33*SPA	-0.656*** (0.094)	0.784	6.173*** (0.703)	2.021	-0.281* (0.129)	0.199	0.974*** (0.274)	0.071	5.175*** (0.866)	1.095
b34	-0.406*** (0.027)	-	-9.178*** (0.409)	-	-3.043*** (0.193)	-	-2.165*** (0.105)	-	-7.426*** (0.418)	-
b34*SPA	0.115 (0.090)	1.571	6.912*** (0.703)	2.776	1.670*** (0.390)	0.578	1.363*** (0.216)	0.722	5.999*** (0.799)	1.849
b35	0.591*** (0.026)	-	-31.565*** (1.495)	-	-6.505*** (0.515)	-	-1.711*** (0.135)	-	-22.452*** (1.649)	-
b35*SPA	-0.379*** (0.092)	1.067	24.528*** (2.557)	2.323	3.805*** (1.023)	0.080	1.255*** (0.273)	0.216	19.508*** (3.316)	1.406
b36	-0.235*** (0.027)	-	-29.494*** (1.359)	-	-37.865*** (2.727)	-	72.985*** (4.211)	-	7.738 (5.928)	-
b36*SPA	0.540*** (0.092)	2.004	23.209*** (2.325)	3.278	22.732*** (5.409)	1.046	-51.459*** (8.178)	1.095	-1.547 (13.811)	2.241
b37	-0.335*** (0.027)	-	-17.891*** (0.816)	-	-9.448*** (0.661)	-	-1.546*** (0.075)	-	-14.237*** (0.930)	-
b37*SPA	0.124 (0.09)	1.580	13.723*** (1.397)	2.812	5.498*** (1.313)	0.594	0.984*** (0.163)	0.732	11.653*** (1.842)	1.896
b38	0.050+ (0.026)	-	-18.973*** (0.884)	-	-14.094*** (1.025)	-	0.050+ (0.026)	-	-15.453*** (1.172)	-
b38*SPA	0.045 (0.091)	1.499	14.779*** (1.513)	2.734	8.385*** (2.035)	0.517	0.045 (0.091)	0.650	12.666*** (2.400)	1.819
b39	-0.063* (0.027)	-	-7.372*** (0.341)	-	-0.062* (0.026)	-	-2.369*** (0.135)	-	-5.766*** (0.378)	-
b39*SPA	-0.536*** (0.090)	0.906	5.135*** (0.587)	2.118	-0.533*** (0.089)	0.085	1.099*** (0.273)	0.057	4.361*** (0.751)	1.190
b40	-0.309*** (0.027)	-	-17.864*** (0.816)	-	-46.827*** (3.371)	-	-0.425*** (0.028)	-	-23.182*** (3.148)	-
b40*SPA	0.219* (0.091)	1.677	13.817*** (1.397)	2.908	27.648*** (6.686)	0.723	0.302** (0.092)	0.829	16.239* (6.956)	2.018

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b41	0.279*** (0.026)	-	-14.358*** (0.680)	-	-4.877*** (0.374)	-	-1.477*** (0.104)	-	-11.055*** (0.724)	-
b41*SPA	-0.434*** (0.090)	1.010	10.906*** (1.166)	2.239	2.608*** (0.746)	0.023	0.812*** (0.216)	0.160	9.102*** (1.414)	1.315
b42	-0.536*** (0.027)	-	-38.583*** (1.767)	-	-72.812*** (5.237)	-	-2.843*** (0.135)	-	-43.856*** (4.826)	-
b42*SPA	0.646*** (0.092)	2.113	30.123*** (3.022)	3.420	43.268*** (10.387)	1.188	2.284*** (0.273)	1.266	32.344** (10.510)	2.571
b43	0.355*** (0.026)	-	-23.053*** (1.088)	-	-15.665*** (1.161)	-	23.662*** (1.341)	-	-6.965** (2.440)	-
b43*SPA	-0.545*** (0.090)	0.897	17.584*** (1.861)	2.140	8.900*** (2.304)	0.085	-17.093*** (2.605)	0.036	7.320 (5.597)	1.200
b44	-3.435*** (0.027)	-	-50.124*** (2.175)	-	-26.685*** (1.686)	-	-2.879*** (0.041)	-	-38.217*** (2.575)	-
b44*SPA	3.552*** (0.094)	5.078	39.685*** (3.720)	6.264	17.259*** (3.343)	4.097	3.157*** (0.112)	4.222	32.772*** (5.228)	5.328
b45	-5.114*** (0.026)	-	-32.940*** (1.292)	-	-25.259*** (1.460)	-	-5.771*** (0.046)	-	-27.995*** (1.674)	-
b45*SPA	3.923*** (0.093)	5.457	25.463*** (2.210)	6.724	15.801*** (2.896)	4.480	4.385*** (0.119)	4.599	22.502*** (3.403)	5.801
delta1	5.360*** (0.010)	-	5.388*** (0.010)	-	5.364*** (0.010)	-	5.360*** (0.010)	-	5.407*** (0.010)	-
delta1*SPA	-1.102*** (0.041)	-	-1.120*** (0.041)	-	-1.107*** (0.041)	-	-1.098*** (0.041)	-	-1.125*** (0.042)	-
delta2	6.979*** (0.014)	-	7.051*** (0.014)	-	6.988*** (0.014)	-	6.986*** (0.014)	-	7.106*** (0.014)	-
delta2*SPA	-0.806*** (0.102)	-	-0.864*** (0.102)	-	-0.816*** (0.102)	-	-0.808*** (0.102)	-	-0.899*** (0.102)	-
delta3	8.323*** (0.019)	-	8.417*** (0.019)	-	8.333*** (0.019)	-	8.334*** (0.019)	-	8.495*** (0.019)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
delta3*SPA	-0.937*** (0.185)	-	-1.016*** (0.186)	-	-0.949*** (0.185)	-	-0.943*** (0.185)	-	-1.072*** (0.186)	-
Intercept Variance	1.039		1.029		1.028		1.02		1.033	
LEX Variance	-		0.029		-		-		0.051	
NP Variance	-		-		0.006		-		0.005	
RC Variance	-		-		-		0.006		0.040	
Intercept*Feature Covariance	-		0.17		0.068		-0.078		See Table G30	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G25.

EPvSPA Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b01*SPA	1.686* (0.090)	[1.510, 1.862]	3.002* (4.183)	[-5.196, 11.201]	0.732* (3.273)	[-5.684, 7.147]	0.803 (8.149)	[-15.169, 16.775]	2.014 (14.994)	[-27.374, 31.402]
b02*SPA	-	-	-	-	-	-	-	-	-	-
b03*SPA	1.042* (0.090)	[0.866, 1.218]	2.263* (1.977)	[-1.612, 6.138]	0.083* (3.148)	[-6.087, 6.254]	0.209* (0.273)	[-0.326, 0.745]	1.371* (3.287)	[-5.072, 7.813]
b04*SPA	1.27 (0.090)	[1.094, 1.446]	2.503* (2.210)	[-1.829, 6.834]	0.304* (1.227)	[-2.101, 2.709]	0.439* (0.273)	[-0.096, 0.975]	1.610* (2.781)	[-3.841, 7.061]
b05*SPA	1.060* (0.091)	[0.882, 1.238]	2.280* (1.977)	[-1.595, 6.155]	0.093* (0.761)	[-1.398, 1.585]	0.226* (0.273)	[-0.309, 0.762]	1.380* (2.526)	[-3.571, 6.331]
b06*SPA	1.101* (0.090)	[0.925, 1.277]	2.292* (0.818)	[0.689, 3.895]	0.144* (3.32)	[-6.363, 6.651]	0.269* (0.273)	[-0.266, 0.805]	1.396* (3.347)	[-5.164, 7.956]
b07*SPA	0.818* (0.089)	[0.644, 0.992]	2.015* (0.933)	[0.186, 3.844]	-0.149* (0.932)	[-1.976, 1.678]	-0.015* (0.273)	[-0.55, 0.521]	1.110* (1.125)	[-1.095, 3.315]
b08*SPA	1.719* (0.090)	[1.543, 1.895]	2.962* (2.325)	[-1.595, 7.519]	0.750* (0.39)	[-0.014, 1.515]	0.886* (0.246)	[0.404, 1.369]	2.069* (3.211)	[-4.225, 8.362]
b09*SPA	1.520 (0.091)	[1.342, 1.698]	2.802* (3.487)	[-4.033, 9.636]	0.556* (1.934)	[-3.234, 4.347]	0.672* (2.544)	[-4.314, 5.658]	1.885 (7.377)	[-12.574, 16.344]
b10*SPA	1.250 (0.096)	[1.062, 1.438]	2.497* (3.487)	[-4.338, 9.331]	0.283* (0.562)	[-0.819, 1.384]	0.407* (2.573)	[-4.636, 5.450]	1.581 (7.608)	[-13.331, 16.492]
b11*SPA	1.959* (0.092)	[1.779, 2.139]	3.116* (0.360)	[2.411, 3.822]	0.990* (1.209)	[-1.380, 3.359]	1.129* (0.273)	[0.594, 1.665]	2.209* (1.175)	[-0.094, 4.512]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b12*SPA	5.760* (0.095)	[5.574, 5.946]	6.942* (3.487)	[0.107, 13.776]	4.793* (4.146)	[-3.333, 12.919]	4.919* (0.274)	[4.382, 5.456]	6.021* (5.028)	[-3.834, 15.876]
b13*SPA	1.787* (0.091)	[1.609, 1.965]	2.978* (1.049)	[0.922, 5.034]	0.813* (0.390)	[0.049, 1.578]	0.957* (0.273)	[0.422, 1.493]	2.074* (1.259)	[-0.394, 4.541]
b14*SPA	1.393 (0.090)	[1.217, 1.569]	2.626* (2.093)	[-1.477, 6.728]	0.432* (2.218)	[-3.916, 4.779]	0.545* (2.682)	[-4.712, 5.802]	1.696* (5.902)	[-9.872, 13.264]
b15*SPA	1.445 (0.091)	[1.267, 1.623]	2.646* (1.281)	[0.136, 5.157]	0.471* (0.127)	[0.222, 0.720]	0.613* (0.273)	[0.078, 1.149]	1.742* (1.738)	[-1.664, 5.149]
b16*SPA	1.461 (0.092)	[1.281, 1.641]	2.703* (2.441)	[-2.081, 7.487]	0.504* (3.208)	[-5.784, 6.791]	0.630* (0.273)	[0.095, 1.166]	1.823* (3.631)	[-5.294, 8.940]
b17*SPA	0.638* (0.094)	[0.454, 0.822]	1.843* (3.370)	[-4.762, 8.448]	-0.329 (9.856)	[-19.647, 18.989]	-0.196 (7.888)	[-15.657, 15.264]	0.944 (16.596)	[-31.585, 33.472]
b18*SPA	1.361 (0.091)	[1.183, 1.539]	2.541* (0.587)	[1.391, 3.692]	0.400* (2.497)	[-4.494, 5.295]	0.530* (0.273)	[-0.005, 1.066]	1.640* (2.515)	[-3.289, 6.570]
b19*SPA	0.955* (0.090)	[0.779, 1.131]	2.134* (0.089)	[1.959, 2.308]	-0.014* (0.126)	[-0.261, 0.233]	0.121* (0.273)	[-0.414, 0.657]	1.221* (0.351)	[0.533, 1.909]
b20*SPA	0.894* (0.090)	[0.718, 1.070]	2.091* (0.703)	[0.713, 3.469]	-0.073* (0.477)	[-1.008, 0.862]	0.059* (0.273)	[-0.476, 0.595]	1.183* (0.757)	[-0.301, 2.667]
b21*SPA	0.748* (0.091)	[0.570, 0.926]	1.955* (1.166)	[-0.330, 4.241]	-0.200 (10.188)	[-20.168, 19.769]	-0.088* (0.273)	[-0.623, 0.448]	1.063 (10.975)	[-20.448, 22.574]
b22*SPA	0.889* (0.093)	[0.707, 1.071]	2.092* (0.703)	[0.714, 3.470]	-0.077* (0.128)	[-0.328, 0.174]	0.053* (0.274)	[-0.484, 0.590]	1.184* (0.865)	[-0.512, 2.879]
b23*SPA	4.985* (0.095)	[4.799, 5.171]	6.206* (2.557)	[1.194, 11.218]	4.029* (2.511)	[-0.893, 8.950]	4.146* (0.165)	[3.823, 4.470]	5.304* (3.503)	[-1.562, 12.169]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
b24*SPA	1.527 (0.090)	[1.351, 1.703]	2.773* (2.441)	[-2.011, 7.557]	0.582 (5.110)	[-9.433, 10.598]	0.696* (0.163)	[0.377, 1.016]	1.901* (5.206)	[-8.303, 12.105]
b25*SPA	2.149* (0.092)	[1.969, 2.329]	3.338* (1.050)	[1.28, 5.396]	1.179* (1.118)	[-1.012, 3.370]	1.320* (0.273)	[0.785, 1.856]	2.439* (1.316)	[-0.140, 5.019]
b26*SPA	1.223* (0.092)	[1.043, 1.403]	2.407* (0.474)	[1.478, 3.337]	0.253* (0.128)	[0.002, 0.504]	0.390* (0.273)	[-0.145, 0.926]	1.497* (0.535)	[0.448, 2.546]
b27*SPA	1.346 (0.090)	[1.170, 1.522]	2.622* (3.370)	[-3.983, 9.227]	0.379* (1.136)	[-1.847, 2.606]	0.515* (0.216)	[0.091, 0.938]	1.738* (4.543)	[-7.167, 10.642]
b28*SPA	2.355* (0.092)	[2.175, 2.535]	3.580* (1.745)	[0.160, 7.000]	1.425 (6.576)	[-11.464, 14.314]	1.443 (10.704)	[-19.536, 22.423]	2.521 (16.927)	[-30.656, 35.698]
b29*SPA	1.515 (0.090)	[1.339, 1.691]	2.690* (0.474)	[1.761, 3.620]	0.548* (1.206)	[-1.816, 2.912]	0.684* (0.273)	[0.149, 1.220]	1.784* (1.137)	[-0.445, 4.012]
b30*SPA	0.920* (0.090)	[0.744, 1.096]	2.139* (1.861)	[-1.508, 5.787]	-0.046* (1.206)	[-2.410, 2.318]	0.087* (0.273)	[-0.448, 0.623]	1.240* (2.304)	[-3.276, 5.756]
b31*SPA	1.430 (0.090)	[1.254, 1.606]	2.641* (1.513)	[-0.325, 5.606]	0.458* (0.390)	[-0.306, 1.223]	0.599* (0.273)	[0.064, 1.135]	1.741* (1.954)	[-2.089, 5.571]
b32*SPA	4.994* (0.095)	[4.808, 5.180]	6.228* (2.210)	[1.896, 10.559]	4.034* (3.044)	[-1.932, 10.001]	4.152* (0.274)	[3.615, 4.689]	5.319* (3.358)	[-1.263, 11.900]
b33*SPA	0.768* (0.094)	[0.584, 0.952]	1.981* (0.703)	[0.603, 3.359]	-0.195* (0.129)	[-0.448, 0.058]	-0.070* (0.274)	[-0.607, 0.467]	1.073* (0.866)	[-0.625, 2.770]
b34*SPA	1.539 (0.090)	[1.363, 1.715]	2.720* (0.703)	[1.342, 4.098]	0.566* (0.390)	[-0.198, 1.331]	0.708* (0.216)	[0.284, 1.131]	1.812* (0.799)	[0.246, 3.378]
b35*SPA	1.045* (0.092)	[0.865, 1.225]	2.276* (2.557)	[-2.736, 7.288]	0.079* (1.023)	[-1.926, 2.084]	0.211* (0.273)	[-0.324, 0.747]	1.378* (3.316)	[-5.122, 7.877]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b36*SPA	1.964* (0.092)	[1.784, 2.144]	3.212* (2.325)	[-1.345, 7.769]	1.025 (5.409)	[-9.577, 11.627]	1.073 (8.178)	[-14.956, 17.102]	2.196 (13.811)	[-24.874, 29.265]
b37*SPA	1.548 (0.090)	[1.372, 1.724]	2.756* (1.397)	[0.018, 5.494]	0.582* (1.313)	[-1.992, 3.155]	0.717* (0.163)	[0.398, 1.037]	1.857* (1.842)	[-1.753, 5.468]
b38*SPA	1.469 (0.091)	[1.291, 1.647]	2.679* (1.513)	[-0.287, 5.644]	0.506* (2.035)	[-3.482, 4.495]	0.637* (0.091)	[0.458, 0.815]	1.782* (2.400)	[-2.922, 6.486]
b39*SPA	0.888* (0.090)	[0.712, 1.064]	2.075* (0.587)	[0.925, 3.226]	-0.083* (0.089)	[-0.257, 0.092]	0.055* (0.273)	[-0.48, 0.591]	1.166* (0.751)	[-0.306, 2.638]
b40*SPA	1.643* (0.091)	[1.465, 1.821]	2.850* (1.397)	[0.112, 5.588]	0.708 (6.686)	[-12.396, 13.813]	0.812* (0.092)	[0.632, 0.992]	1.977 (6.956)	[-11.656, 15.611]
b41*SPA	0.990* (0.090)	[0.814, 1.166]	2.193* (1.166)	[-0.092, 4.479]	0.023* (0.746)	[-1.439, 1.485]	0.157* (0.216)	[-0.267, 0.580]	1.288* (1.414)	[-1.483, 4.060]
b42*SPA	2.070* (0.092)	[1.890, 2.250]	3.351* (3.022)	[-2.572, 9.274]	1.164 (10.387)	[-19.195, 21.523]	1.240* (0.273)	[0.705, 1.776]	2.519 (10.51)	[-18.081, 23.119]
b43*SPA	0.879* (0.090)	[0.703, 1.055]	2.096* (1.861)	[-1.551, 5.744]	-0.083* (2.304)	[-4.599, 4.432]	0.035* (2.605)	[-5.07, 5.141]	1.176* (5.597)	[-9.794, 12.146]
b44*SPA	4.976* (0.094)	[4.792, 5.160]	6.138* (3.720)	[-1.154, 13.429]	4.014* (3.343)	[-2.538, 10.566]	4.137* (0.112)	[3.918, 4.357]	5.221* (5.228)	[-5.026, 15.467]
b45*SPA	5.347* (0.093)	[5.165, 5.529]	6.588* (2.210)	[2.256, 10.919]	4.389* (2.896)	[-1.287, 10.066]	4.507* (0.119)	[4.273, 4.740]	5.684* (3.403)	[-0.986, 12.354]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G26.

EPvOTH Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	-0.999*** (0.021)	-	21.173*** (1.042)	-	15.809*** (1.236)	-	-8.511*** (0.439)	-	13.433*** (1.670)	-
Intercept*OTH	1.112*** (0.090)	-	-14.419*** (2.150)	-	-10.449*** (3.068)	-	4.733*** (1.150)	-	-11.738* (4.691)	-
LEX	-	-	13.681*** (0.644)	-	-	-	-	-	8.546*** (0.801)	-
LEX*OTH	-	-	-9.553*** (1.333)	-	-	-	-	-	-6.327*** (1.744)	-
NP	-	-	-	-	20.585*** (1.515)	-	-	-	5.175*** (1.530)	-
NP*OTH	-	-	-	-	-14.139*** (3.765)	-	-	-	-5.175 (3.764)	-
RC	-	-	-	-	-	-	-28.820*** (1.682)	-	-14.07*** (1.774)	-
RC*OTH	-	-	-	-	-	-	14.034** (4.361)	-	6.535 (5.266)	-
b01	-0.370*** (0.027)	-	-53.033*** (2.475)	-	-23.111*** (1.673)	-	72.645*** (4.258)	-	-3.345 (6.496)	-
b01*OTH	0.265* (0.114)	1.405	37.091*** (5.118)	2.616	15.891*** (4.158)	0.287	-35.314** (11.043)	0.758	13.803 (18.861)	1.474
b02	-	-	-	-	-	-	-	-	-	-
b02*OTH	-	-	-	-	-	-	-	-	-	-
b03	0.344*** (0.026)	-	-24.508*** (1.169)	-	-21.528*** (1.609)	-	-1.961*** (0.137)	-	-21.838*** (1.632)	-

Effect	Base model Estimate (SE)	Base model Effect Size	LEX Predictor Estimate (SE)	LEX Predictor Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size	RC Predictor Estimate (SE)	RC Predictor Effect Size	All Predictors Estimate (SE)	All Predictors Effect Size
b03*OTH	-0.216+ (0.114)	0.914	17.161*** (2.418)	2.038	14.813*** (4.000)	0.207	0.907* (0.367)	0.313	17.352*** (4.334)	0.986
b04	0.079** (0.026)	-	-27.710*** (1.307)	-	-8.427*** (0.626)	-	-2.227*** (0.137)	-	-20.584*** (1.418)	-
b04*OTH	0.048 (0.114)	1.184	19.480*** (2.703)	2.318	5.893*** (1.559)	0.054	1.172** (0.367)	0.584	15.624*** (3.295)	1.268
b05	0.478*** (0.026)	-	-24.372*** (1.169)	-	-4.777*** (0.387)	-	-1.826*** (0.137)	-	-17.519*** (1.277)	-
b05*OTH	-0.233* (0.115)	0.897	17.143*** (2.418)	2.020	3.378*** (0.967)	0.232	0.890* (0.367)	0.296	13.152*** (2.854)	0.961
b06	-0.077** (0.027)	-	-10.316*** (0.482)	-	-23.146*** (1.697)	-	-2.384*** (0.137)	-	-13.422*** (1.545)	-
b06*OTH	-0.324** (0.114)	0.804	6.842*** (1.002)	1.900	15.526*** (4.218)	0.316	0.799* (0.367)	0.203	10.768** (4.007)	0.843
b07	0.060* (0.026)	-	-11.635*** (0.550)	-	-6.387*** (0.475)	-	-2.247*** (0.137)	-	-10.011*** (0.597)	-
b07*OTH	-0.493*** (0.113)	0.632	7.690*** (1.142)	1.732	3.939*** (1.184)	0.497	0.629+ (0.367)	0.030	7.091*** (1.464)	0.668
b08	-0.953*** (0.028)	-	-30.210*** (1.375)	-	-3.589*** (0.196)	-	-3.001*** (0.123)	-	-20.942*** (1.582)	-
b08*OTH	0.122 (0.115)	1.259	20.572*** (2.845)	2.399	1.938*** (0.495)	0.131	1.118*** (0.330)	0.658	14.836*** (3.433)	1.344
b09	-0.005 (0.026)	-	-43.877*** (2.062)	-	-13.435*** (0.988)	-	22.773*** (1.329)	-	-19.706*** (3.321)	-
b09*OTH	0.292* (0.115)	1.433	30.985*** (4.266)	2.624	9.519*** (2.457)	0.306	-10.809** (3.447)	0.817	18.870* (8.958)	1.547
b10	0.911*** (0.026)	-	-42.899*** (2.062)	-	-2.949*** (0.284)	-	23.941*** (1.344)	-	-16.203*** (3.397)	-
b10*OTH	-0.185 (0.118)	0.946	30.471*** (4.266)	2.099	2.465*** (0.714)	0.183	-11.407** (3.487)	0.336	15.892+ (8.742)	1.023

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b11	-0.296*** (0.027)	-	-4.677*** (0.208)	-	-8.679*** (0.617)	-	-2.604*** (0.137)	-	-6.274*** (0.550)	-
b11*OTH	0.304** (0.114)	1.445	3.364*** (0.441)	2.513	6.064*** (1.537)	0.316	1.429*** (0.367)	0.846	4.971*** (1.401)	1.450
b12	-3.940*** (0.026)	-	-47.714*** (2.063)	-	-32.740*** (2.120)	-	-6.236*** (0.137)	-	-39.629*** (2.533)	-
b12*OTH	3.936*** (0.116)	5.152	34.486*** (4.266)	6.197	23.715*** (5.268)	4.015	5.048*** (0.367)	4.540	31.895*** (6.593)	5.092
b13	-0.304*** (0.027)	-	-13.463*** (0.619)	-	-2.938*** (0.196)	-	-2.611*** (0.137)	-	-10.333*** (0.652)	-
b13*OTH	0.122 (0.114)	1.259	9.324*** (1.284)	2.356	1.933*** (0.495)	0.125	1.246*** (0.367)	0.659	7.424*** (1.41)	1.294
b14	-0.268*** (0.027)	-	-26.597*** (1.238)	-	-15.675*** (1.134)	-	23.752*** (1.401)	-	-8.876*** (2.605)	-
b14*OTH	-0.273* (0.114)	0.856	18.132*** (2.561)	1.986	10.314*** (2.818)	0.268	-11.976*** (3.635)	0.242	10.356 (7.537)	0.902
b15	0.012 (0.026)	-	-16.070*** (0.756)	-	-0.605*** (0.053)	-	-2.294*** (0.137)	-	-11.338*** (0.864)	-
b15*OTH	-0.025 (0.114)	1.109	11.219*** (1.567)	2.213	0.400* (0.160)	0.025	1.099** (0.367)	0.509	8.125*** (1.807)	1.152
b16	0.163*** (0.026)	-	-30.544*** (1.444)	-	-22.122*** (1.639)	-	-2.143*** (0.137)	-	-25.791*** (1.828)	-
b16*OTH	-0.001 (0.115)	1.134	21.472*** (2.987)	2.275	15.312*** (4.075)	0.014	1.123** (0.367)	0.534	20.399*** (4.790)	1.232
b17	1.339*** (0.026)	-	-40.977*** (1.993)	-	-67.109*** (5.039)	-	71.958*** (4.122)	-	-7.866 (7.319)	-
b17*OTH	-0.454*** (0.120)	0.672	29.155*** (4.123)	1.799	46.589*** (12.523)	0.447	-34.864** (10.691)	0.057	20.370 (21.739)	0.724
b18	0.114*** (0.026)	-	-7.197*** (0.344)	-	-17.229*** (1.276)	-	-2.192*** (0.137)	-	-9.956*** (1.160)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b18*OTH	-0.120 (0.114)	1.012	4.992*** (0.719)	2.098	11.797*** (3.172)	0.111	1.003** (0.367)	0.411	8.168** (2.982)	1.040
b19	0.351*** (0.026)	-	0.334*** (0.025)	-	-0.270*** (0.052)	-	-1.954*** (0.137)	-	-0.951*** (0.149)	-
b19*OTH	-0.405*** (0.114)	0.722	-0.388*** (0.112)	1.804	0.023 (0.160)	0.410	0.717+ (0.367)	0.119	0.296 (0.427)	0.736
b20	0.518*** (0.026)	-	-8.271*** (0.413)	-	-2.740*** (0.241)	-	-1.786*** (0.137)	-	-6.936*** (0.413)	-
b20*OTH	-0.337** (0.115)	0.791	5.807*** (0.860)	1.887	1.902** (0.605)	0.339	0.784* (0.367)	0.188	5.093*** (0.931)	0.823
b21	0.681*** (0.026)	-	-13.958*** (0.688)	-	-70.104*** (5.209)	-	-1.622*** (0.137)	-	-27.408*** (4.963)	-
b21*OTH	-0.618*** (0.114)	0.504	9.619*** (1.426)	1.614	48.011*** (12.944)	0.611	0.502 (0.367)	0.100	24.510 (12.576)	0.559
b22	0.888*** (0.026)	-	-7.914*** (0.413)	-	0.260*** (0.052)	-	-1.413*** (0.137)	-	-5.917*** (0.445)	-
b22*OTH	-0.520*** (0.116)	0.604	5.634*** (0.860)	1.710	-0.088 (0.161)	0.523	0.601 (0.368)	0.001	4.256*** (0.910)	0.645
b23	-3.319*** (0.027)	-	-35.457*** (1.512)	-	-20.768*** (1.284)	-	-4.520*** (0.076)	-	-28.392*** (1.769)	-
b23*OTH	2.773*** (0.117)	3.965	25.22*** (3.128)	5.066	14.764*** (3.191)	2.845	3.353*** (0.217)	3.354	22.306*** (4.507)	3.989
b24	-0.228*** (0.027)	-	-30.945*** (1.444)	-	-35.74*** (2.612)	-	-1.440*** (0.076)	-	-28.992*** (2.501)	-
b24*OTH	0.031 (0.114)	1.167	21.505*** (2.987)	2.308	24.430*** (6.492)	0.055	0.621** (0.216)	0.566	23.511*** (6.748)	1.271
b25	-0.475*** (0.027)	-	-13.635*** (0.619)	-	-8.219*** (0.570)	-	-2.783*** (0.137)	-	-11.792*** (0.691)	-
b25*OTH	0.471*** (0.114)	1.616	9.670*** (1.284)	2.709	5.791*** (1.420)	0.484	1.596*** (0.367)	1.017	9.058*** (1.722)	1.652

Effect	Base model Estimate (SE)	Base model Effect Size	LEX Predictor Estimate (SE)	LEX Predictor Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size	RC Predictor Estimate (SE)	RC Predictor Effect Size	All Predictors Estimate (SE)	All Predictors Effect Size
b26	0.464*** (0.026)	-	-5.407*** (0.276)	-	-0.159** (0.052)	-	-1.840*** (0.137)	-	-4.502*** (0.285)	-
b26*OTH	-0.295* (0.114)	0.834	3.809*** (0.580)	1.924	0.133 (0.160)	0.297	0.827* (0.367)	0.232	3.117*** (0.580)	0.858
b27	-0.378*** (0.027)	-	-42.798*** (1.993)	-	-8.245*** (0.579)	-	-2.138*** (0.106)	-	-29.780*** (2.254)	-
b27*OTH	-0.161 (0.114)	0.971	29.485*** (4.123)	2.136	5.246*** (1.443)	0.159	0.695* (0.290)	0.369	21.908*** (5.089)	1.088
b28	-0.673*** (0.028)	-	-22.616*** (1.032)	-	-46.385*** (3.362)	-	95.240*** (5.594)	-	21.017** (7.345)	-
b28*OTH	0.825*** (0.115)	1.977	16.169*** (2.135)	3.102	32.238*** (8.355)	0.881	-45.922** (14.506)	1.304	0.629 (22.180)	1.899
b29	-0.492*** (0.027)	-	-6.339*** (0.277)	-	-8.854*** (0.616)	-	-2.800*** (0.137)	-	-7.383*** (0.546)	-
b29*OTH	-0.112 (0.114)	1.021	3.986*** (0.580)	2.105	5.636*** (1.533)	0.107	1.011** (0.367)	0.420	5.242*** (1.409)	1.041
b30	0.198*** (0.026)	-	-23.203*** (1.100)	-	-8.165*** (0.616)	-	-2.107*** (0.137)	-	-17.68*** (1.186)	-
b30*OTH	-0.639*** (0.113)	0.483	15.719*** (2.278)	1.600	5.110*** (1.533)	0.644	0.482 (0.367)	0.120	12.849*** (2.794)	0.539
b31	-0.106*** (0.027)	-	-19.122*** (0.894)	-	-2.741*** (0.196)	-	-2.413*** (0.137)	-	-13.802*** (0.985)	-
b31*OTH	-0.244* (0.114)	0.886	13.052*** (1.852)	1.998	1.569** (0.495)	0.246	0.878* (0.367)	0.284	9.776*** (2.127)	0.937
b32	-3.306*** (0.027)	-	-31.097*** (1.307)	-	-24.456*** (1.556)	-	-5.600*** (0.137)	-	-27.132*** (1.687)	-
b32*OTH	2.842*** (0.116)	4.035	22.26*** (2.703)	5.156	17.372*** (3.868)	2.910	3.951*** (0.367)	3.420	21.545*** (4.435)	4.068
b33	1.250*** (0.026)	-	-7.567*** (0.413)	-	0.617*** (0.052)	-	-1.048*** (0.137)	-	-5.571*** (0.445)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b33*OTH	-0.500*** (0.119)	0.625	5.658*** (0.861)	1.735	-0.068 (0.163)	0.503	0.619+ (0.369)	0.019	4.280*** (0.910)	0.669
b34	-0.406*** (0.027)	-	-9.175*** (0.413)	-	-3.040*** (0.196)	-	-2.166*** (0.106)	-	-7.416*** (0.423)	-
b34*OTH	0.362** (0.114)	1.504	6.492*** (0.860)	2.586	2.172*** (0.495)	0.369	1.220*** (0.290)	0.905	5.496*** (0.945)	1.520
b35	0.592*** (0.026)	-	-31.553*** (1.512)	-	-6.497*** (0.522)	-	-1.712*** (0.137)	-	-22.422*** (1.668)	-
b35*OTH	-0.411*** (0.115)	0.715	22.065*** (3.128)	1.846	4.460*** (1.300)	0.412	0.711+ (0.367)	0.113	16.815*** (3.769)	0.788
b36	-0.235*** (0.027)	-	-29.484*** (1.375)	-	-37.828*** (2.765)	-	73.037*** (4.273)	-	7.842 (5.994)	-
b36*OTH	0.448*** (0.115)	1.592	20.905*** (2.845)	2.739	26.282*** (6.872)	0.488	-35.264** (11.083)	0.938	6.773 (18.044)	1.573
b37	-0.336*** (0.027)	-	-17.885*** (0.825)	-	-9.439*** (0.670)	-	-1.548*** (0.076)	-	-14.207*** (0.940)	-
b37*OTH	0.153 (0.114)	1.291	12.425*** (1.710)	2.401	6.409*** (1.668)	0.163	0.743*** (0.216)	0.690	10.874*** (2.361)	1.344
b38	0.050+ (0.026)	-	-18.967*** (0.894)	-	-14.08*** (1.040)	-	0.050+ (0.026)	-	-15.410*** (1.185)	-
b38*OTH	-0.097 (0.114)	1.036	13.199*** (1.852)	2.148	9.612*** (2.585)	0.089	-0.098 (0.114)	0.434	12.290*** (3.161)	1.090
b39	-0.063* (0.027)	-	-7.369*** (0.345)	-	-0.062* (0.026)	-	-2.370*** (0.137)	-	-5.763*** (0.383)	-
b39*OTH	-0.481*** (0.114)	0.644	4.637*** (0.719)	1.736	-0.477*** (0.113)	0.487	0.641+ (0.367)	0.042	3.447*** (0.772)	0.669
b40	-0.309*** (0.027)	-	-17.858*** (0.825)	-	-46.781*** (3.418)	-	-0.425*** (0.028)	-	-23.054*** (3.183)	-
b40*OTH	0.201+ (0.114)	1.340	12.472*** (1.710)	2.449	32.132*** (8.494)	0.239	0.258* (0.115)	0.740	20.081* (8.323)	1.413

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b41	0.279*** (0.026)	-	-14.353*** (0.688)	-	-4.871*** (0.380)	-	-1.478*** (0.106)	-	-11.036*** (0.733)	-
b41*OTH	-0.292* (0.114)	0.837	9.940*** (1.426)	1.942	3.249*** (0.948)	0.292	0.564+ (0.290)	0.235	8.201*** (1.711)	0.879
b42	-0.537*** (0.027)	-	-38.570*** (1.787)	-	-72.740*** (5.310)	-	-2.845*** (0.137)	-	-43.654*** (4.880)	-
b42*OTH	0.378*** (0.114)	1.521	26.972*** (3.697)	2.691	49.989*** (13.197)	0.440	1.503*** (0.367)	0.922	36.753** (12.925)	1.684
b43	0.355*** (0.026)	-	-23.045*** (1.100)	-	-15.648*** (1.177)	-	23.679*** (1.361)	-	-6.916** (2.467)	-
b43*OTH	-0.514*** (0.114)	0.610	15.844*** (2.278)	1.727	10.482*** (2.927)	0.515	-11.878*** (3.530)	0.002	9.069 (7.209)	0.651
b44	-3.437*** (0.027)	-	-50.109*** (2.200)	-	-26.664*** (1.709)	-	-2.881*** (0.042)	-	-38.142*** (2.604)	-
b44*OTH	3.058*** (0.116)	4.256	35.618*** (4.551)	5.276	19.013*** (4.248)	3.127	2.783*** (0.143)	3.646	30.298*** (6.640)	4.178
b45	-5.117*** (0.026)	-	-32.933*** (1.307)	-	-25.242*** (1.480)	-	-5.775*** (0.047)	-	-27.936*** (1.693)	-
b45*OTH	2.457*** (0.123)	3.642	21.908*** (2.704)	4.796	16.287*** (3.680)	2.524	2.775*** (0.159)	3.036	20.599*** (4.481)	3.747
delta1	5.364*** (0.010)	-	5.392*** (0.010)	-	5.368*** (0.010)	-	5.365*** (0.010)	-	5.412*** (0.010)	-
delta1*OTH	-1.111*** (0.043)	-	-1.112*** (0.044)	-	-1.115*** (0.043)	-	-1.104*** (0.043)	-	-1.101*** (0.044)	-
delta2	6.985*** (0.014)	-	7.057*** (0.014)	-	6.993*** (0.014)	-	6.992*** (0.014)	-	7.112*** (0.014)	-
delta2*OTH	-0.943*** (0.091)	-	-0.971*** (0.091)	-	-0.949*** (0.091)	-	-0.939*** (0.091)	-	-0.979*** (0.092)	-
delta3	8.329*** (0.019)	-	8.424*** (0.019)	-	8.340*** (0.019)	-	8.34*** (0.019)	-	8.502*** (0.019)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
delta3*OTH	-1.103*** (0.158)	-	-1.148*** (0.159)	-	-1.112*** (0.158)	-	-1.103*** (0.158)	-	-1.170*** (0.160)	-
Intercept Variance	1.074		1.063		1.063		1.054		1.068	
LEX Variance	-		0.029		-		-		0.053	
NP Variance	-		-		0.006		-		0.006	
RC Variance	-		-		-		0.006		0.040	
Intercept*Feature Covariance	-		0.176		0.072		-0.079		See Table G30	

Note: + p < .10, * p < .05, ** p < .01, *** p < .001. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G27.

EPvOTH Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b01*OTH	1.377* (0.114)	[1.154, 1.600]	2.563* (5.118)	[-7.468, 12.594]	0.281* (4.158)	[-7.868, 8.431]	0.743 (11.043)	[-20.901, 22.387]	1.444 (18.861)	[-35.524, 38.411]
b02*OTH	-	-	-	-	-	-	-	-	-	-
b03*OTH	0.896 (0.114)	[0.673, 1.119]	1.997* (2.418)	[-2.742, 6.736]	-0.203* (4.000)	[-8.043, 7.637]	0.307* (0.367)	[-0.412, 1.026]	0.966* (4.334)	[-7.529, 9.460]
b04*OTH	1.160 (0.114)	[0.937, 1.383]	2.272* (2.703)	[-3.026, 7.569]	0.053* (1.559)	[-3.002, 3.109]	0.572* (0.367)	[-0.147, 1.291]	1.242* (3.295)	[-5.216, 7.700]
b05*OTH	0.879* (0.115)	[0.654, 1.104]	1.979* (2.418)	[-2.760, 6.718]	-0.228* (0.967)	[-2.123, 1.668]	0.290* (0.367)	[-0.429, 1.009]	0.942* (2.854)	[-4.652, 6.536]
b06*OTH	0.788* (0.114)	[0.565, 1.011]	1.861* (1.002)	[-0.103, 3.825]	-0.310* (4.218)	[-8.577, 7.957]	0.199* (0.367)	[-0.520, 0.918]	0.826* (4.007)	[-7.028, 8.680]
b07*OTH	0.619* (0.113)	[0.398, 0.840]	1.697* (1.142)	[-0.542, 3.935]	-0.487* (1.184)	[-2.807, 1.834]	0.029* (0.367)	[-0.69, 0.748]	0.655* (1.464)	[-2.215, 3.524]
b08*OTH	1.234 (0.115)	[1.009, 1.459]	2.351* (2.845)	[-3.225, 7.927]	0.128* (0.495)	[-0.842, 1.098]	0.644* (0.330)	[-0.002, 1.291]	1.317* (3.433)	[-5.411, 8.046]
b09*OTH	1.404* (0.115)	[1.179, 1.629]	2.571* (4.266)	[-5.791, 10.932]	0.300* (2.457)	[-4.516, 5.116]	0.801 (3.447)	[-5.955, 7.557]	1.515 (8.958)	[-16.042, 19.073]
b10*OTH	0.927 (0.118)	[0.696, 1.158]	2.057* (4.266)	[-6.305, 10.418]	-0.179* (0.714)	[-1.579, 1.220]	0.329 (3.487)	[-6.506, 7.163]	1.003 (8.742)	[-16.132, 18.137]
b11*OTH	1.416* (0.114)	[1.193, 1.639]	2.462* (0.441)	[1.598, 3.327]	0.309* (1.537)	[-2.703, 3.322]	0.829* (0.367)	[0.110, 1.548]	1.421* (1.401)	[-1.325, 4.166]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
b12*OTH	5.048* (0.116)	[4.821, 5.275]	6.072* (4.266)	[-2.290, 14.433]	3.934* (5.268)	[-6.391, 14.260]	4.448 (0.367)	[3.729, 5.167]	4.989* (6.593)	[-7.933, 17.911]
b13*OTH	1.234 (0.114)	[1.011, 1.457]	2.309* (1.284)	[-0.208, 4.825]	0.123* (0.495)	[-0.847, 1.093]	0.646* (0.367)	[-0.073, 1.365]	1.268* (1.410)	[-1.496, 4.032]
b14*OTH	0.839* (0.114)	[0.616, 1.062]	1.946* (2.561)	[-3.074, 6.965]	-0.262* (2.818)	[-5.786, 5.261]	0.237 (3.635)	[-6.887, 7.362]	0.884 (7.537)	[-13.888, 15.657]
b15*OTH	1.087 (0.114)	[0.864, 1.310]	2.169* (1.567)	[-0.903, 5.240]	-0.024* (0.160)	[-0.338, 0.289]	0.499* (0.367)	[-0.220, 1.218]	1.129* (1.807)	[-2.413, 4.670]
b16*OTH	1.111 (0.115)	[0.886, 1.336]	2.229* (2.987)	[-3.626, 8.083]	0.013* (4.075)	[-7.974, 8.000]	0.523* (0.367)	[-0.196, 1.242]	1.208* (4.790)	[-8.181, 10.596]
b17*OTH	0.658* (0.120)	[0.423, 0.893]	1.763* (4.123)	[-6.318, 9.844]	-0.438 (12.523)	[-24.983, 24.107]	0.056 (10.691)	[-20.898, 21.010]	0.709 (21.739)	[-41.899, 43.317]
b18*OTH	0.992 (0.114)	[0.769, 1.215]	2.056* (0.719)	[0.646, 3.465]	-0.108* (3.172)	[-6.325, 6.109]	0.403* (0.367)	[-0.316, 1.122]	1.019* (2.982)	[-4.826, 6.863]
b19*OTH	0.707* (0.114)	[0.484, 0.930]	1.767* (0.112)	[1.548, 1.987]	-0.401* (0.160)	[-0.715, -0.088]	0.117* (0.367)	[-0.602, 0.836]	0.721* (0.427)	[-0.116, 1.558]
b20*OTH	0.775* (0.115)	[0.55, 1.000]	1.849* (0.860)	[0.163, 3.534]	-0.332* (0.605)	[-1.518, 0.854]	0.184* (0.367)	[-0.535, 0.903]	0.806* (0.931)	[-1.018, 2.631]
b21*OTH	0.494* (0.114)	[0.271, 0.717]	1.581* (1.426)	[-1.214, 4.376]	-0.599 (12.944)	[-25.969, 24.771]	-0.098* (0.367)	[-0.817, 0.621]	0.548 (12.576)	[-24.101, 25.197]
b22*OTH	0.592* (0.116)	[0.365, 0.819]	1.676* (0.860)	[-0.010, 3.361]	-0.512* (0.161)	[-0.828, -0.197]	0.001* (0.368)	[-0.720, 0.722]	0.632* (0.910)	[-1.152, 2.415]
b23*OTH	3.885* (0.117)	[3.656, 4.114]	4.964* (3.128)	[-1.167, 11.095]	2.788* (3.191)	[-3.466, 9.042]	3.286* (0.217)	[2.861, 3.712]	3.908* (4.507)	[-4.925, 12.742]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b24*OTH	1.143 (0.114)	[0.920, 1.366]	2.262* (2.987)	[-3.593, 8.116]	0.054 (6.492)	[-12.670, 12.778]	0.554* (0.216)	[0.131, 0.978]	1.246 (6.748)	[-11.981, 14.472]
b25*OTH	1.583* (0.114)	[1.360, 1.806]	2.655* (1.284)	[0.138, 5.171]	0.474* (1.420)	[-2.309, 3.258]	0.996* (0.367)	[0.277, 1.715]	1.619* (1.722)	[-1.756, 4.994]
b26*OTH	0.817* (0.114)	[0.594, 1.040]	1.885* (0.580)	[0.749, 3.022]	-0.291* (0.160)	[-0.605, 0.022]	0.227* (0.367)	[-0.492, 0.946]	0.840* (0.580)	[-0.296, 1.977]
b27*OTH	0.951 (0.114)	[0.728, 1.174]	2.093* (4.123)	[-5.988, 10.174]	-0.155* (1.443)	[-2.984, 2.673]	0.362* (0.290)	[-0.207, 0.930]	1.066* (5.089)	[-8.908, 11.041]
b28*OTH	1.937* (0.115)	[1.712, 2.162]	3.040* (2.135)	[-1.145, 7.224]	0.863 (8.355)	[-15.513, 17.239]	1.278 (14.506)	[-27.154, 29.710]	1.861 (22.180)	[-41.612, 45.334]
b29*OTH	1.000 (0.114)	[0.777, 1.223]	2.062* (0.580)	[0.926, 3.199]	-0.105* (1.533)	[-3.109, 2.900]	0.411* (0.367)	[-0.308, 1.130]	1.020* (1.409)	[-1.742, 3.781]
b30*OTH	0.473* (0.113)	[0.252, 0.694]	1.567* (2.278)	[-2.897, 6.032]	-0.631* (1.533)	[-3.635, 2.374]	-0.118* (0.367)	[-0.837, 0.601]	0.528* (2.794)	[-4.948, 6.004]
b31*OTH	0.868* (0.114)	[0.645, 1.091]	1.957* (1.852)	[-1.672, 5.587]	-0.241* (0.495)	[-1.211, 0.729]	0.278* (0.367)	[-0.441, 0.997]	0.918* (2.127)	[-3.25, 5.087]
b32*OTH	3.954* (0.116)	[3.727, 4.181]	5.052* (2.703)	[-0.246, 10.349]	2.851* (3.868)	[-4.730, 10.432]	3.351* (0.367)	[2.632, 4.070]	3.986* (4.435)	[-4.707, 12.678]
b33*OTH	0.612* (0.119)	[0.379, 0.845]	1.700* (0.861)	[0.012, 3.387]	-0.492* (0.163)	[-0.812, -0.173]	0.019* (0.369)	[-0.704, 0.742]	0.656* (0.910)	[-1.128, 2.439]
b34*OTH	1.474* (0.114)	[1.251, 1.697]	2.534* (0.860)	[0.848, 4.219]	0.362* (0.495)	[-0.608, 1.332]	0.887* (0.290)	[0.318, 1.455]	1.489* (0.945)	[-0.363, 3.341]
b35*OTH	0.701* (0.115)	[0.476, 0.926]	1.809* (3.128)	[-4.322, 7.940]	-0.404* (1.300)	[-2.952, 2.144]	0.111* (0.367)	[-0.608, 0.830]	0.772* (3.769)	[-6.615, 8.159]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b36*OTH	1.560* (0.115)	[1.335, 1.785]	2.684* (2.845)	[-2.892, 8.260]	0.478 (6.872)	[-12.991, 13.947]	0.919 (11.083)	[-20.803, 22.642]	1.542 (18.044)	[-33.825, 36.908]
b37*OTH	1.265 (0.114)	[1.042, 1.488]	2.353* (1.710)	[-0.999, 5.704]	0.159* (1.668)	[-3.110, 3.429]	0.676* (0.216)	[0.253, 1.100]	1.317* (2.361)	[-3.311, 5.944]
b38*OTH	1.015 (0.114)	[0.792, 1.238]	2.104* (1.852)	[-1.525, 5.734]	-0.088* (2.585)	[-5.154, 4.979]	0.425* (0.114)	[0.201, 0.648]	1.068* (3.161)	[-5.128, 7.263]
b39*OTH	0.631* (0.114)	[0.408, 0.854]	1.701* (0.719)	[0.291, 3.110]	-0.477* (0.113)	[-0.699, -0.256]	0.041* (0.367)	[-0.678, 0.760]	0.655* (0.772)	[-0.858, 2.168]
b40*OTH	1.313 (0.114)	[1.090, 1.536]	2.400* (1.710)	[-0.952, 5.751]	0.234 (8.494)	[-16.414, 16.882]	0.725* (0.115)	[0.499, 0.950]	1.385 (8.323)	[-14.928, 17.698]
b41*OTH	0.820* (0.114)	[0.597, 1.043]	1.902* (1.426)	[-0.893, 4.697]	-0.286* (0.948)	[-2.144, 1.572]	0.231* (0.290)	[-0.338, 0.799]	0.861* (1.711)	[-2.493, 4.214]
b42*OTH	1.490* (0.114)	[1.267, 1.713]	2.637* (3.697)	[-4.609, 9.883]	0.432 (13.197)	[-25.435, 26.298]	0.903* (0.367)	[0.184, 1.622]	1.650 (12.925)	[-23.683, 26.983]
b43*OTH	0.598* (0.114)	[0.375, 0.821]	1.692* (2.278)	[-2.772, 6.157]	-0.504* (2.927)	[-6.241, 5.233]	-0.002 (3.530)	[-6.920, 6.917]	0.638 (7.209)	[-13.492, 14.767]
b44*OTH	4.170* (0.116)	[3.943, 4.397]	5.169* (4.551)	[-3.751, 14.089]	3.064* (4.248)	[-5.262, 11.390]	3.572* (0.143)	[3.292, 3.853]	4.094* (6.640)	[-8.921, 17.108]
b45*OTH	3.569* (0.123)	[3.328, 3.810]	4.700* (2.704)	[-0.600, 9.999]	2.473* (3.680)	[-4.740, 9.686]	2.975* (0.159)	[2.663, 3.287]	3.671* (4.481)	[-5.112, 12.454]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G28.

OTHvSPA Model Results – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
Intercept	0.114 (0.082)	-	6.906*** (1.386)	-	5.836** (1.990)	-	-3.895*** (0.745)	-	1.555 (3.266)	-
Intercept*SPA	0.307** (0.104)	-	-1.832 (1.742)	-	1.013 (2.486)	-	1.968* (0.909)	-	0.298 (4.048)	-
LEX	-	-	4.222*** (0.859)	-	-	-	-	-	2.027+ (1.159)	-
LEX*SPA	-	-	-1.320 (1.080)	-	-	-	-	-	-1.860 (1.606)	-
NP	-	-	-	-	7.030** (2.443)	-	-	-	0.710 (2.576)	-
NP*SPA	-	-	-	-	0.879 (3.053)	-	-	-	2.792 (3.458)	-
RC	-	-	-	-	-	-	-15.231*** (2.812)	-	-9.089* (3.697)	-
RC*SPA	-	-	-	-	-	-	6.336+ (3.428)	-	2.773 (4.638)	-
b01	-0.102 (0.109)	-	-16.302*** (3.302)	-	-7.862** (2.699)	-	38.462*** (7.122)	-	14.349 (13.204)	-
b01*SPA	-0.003 (0.139)	0.310	5.084 (4.151)	0.483	-0.973 (3.373)	0.368	-16.052+ (8.682)	0.059	-2.951 (16.677)	0.653
b02	-	-	-	-	-	-	-	-	-	-
b02*SPA	-	-	-	-	-	-	-	-	-	-
b03	0.126 (0.110)	-	-7.520*** (1.562)	-	-7.339** (2.596)	-	-1.092*** (0.251)	-	-5.018+ (2.990)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size	Estimate (SE)	Effect Size
b03*SPA	-0.163 (0.140)	0.147	2.237 (1.964)	0.308	-1.095 (3.245)	0.206	0.343 (0.308)	0.099	0.475 (3.683)	0.486
b04	0.125 (0.110)	-	-8.423*** (1.745)	-	-2.779** (1.015)	-	-1.093*** (0.251)	-	-4.991* (2.218)	-
b04*SPA	-0.198 (0.140)	0.111	2.484 (2.194)	0.272	-0.560 (1.268)	0.170	0.308 (0.308)	0.134	2.649 (2.860)	0.449
b05	0.241* (0.111)	-	-7.405*** (1.562)	-	-1.553* (0.633)	-	-0.977*** (0.251)	-	-4.329* (1.907)	-
b05*SPA	-0.128 (0.140)	0.183	2.271 (1.964)	0.343	-0.352 (0.791)	0.240	0.378 (0.308)	0.063	2.762 (2.519)	0.520
b06	-0.392*** (0.109)	-	-3.536*** (0.651)	-	-8.265** (2.738)	-	-1.612*** (0.250)	-	-3.419 (2.763)	-
b06*SPA	-0.001 (0.138)	0.312	0.988 (0.819)	0.470	-0.985 (3.422)	0.370	0.506 (0.307)	0.068	-1.512 (3.557)	0.647
b07	-0.423*** (0.109)	-	-4.015*** (0.741)	-	-2.622*** (0.772)	-	-1.644*** (0.250)	-	-3.094** (0.997)	-
b07*SPA	-0.115 (0.138)	0.196	1.015 (0.932)	0.354	-0.390 (0.966)	0.254	0.392 (0.307)	0.049	0.826 (1.230)	0.532
b08	-0.815*** (0.110)	-	-9.823*** (1.836)	-	-1.713*** (0.332)	-	-1.899*** (0.228)	-	-5.884* (2.280)	-
b08*SPA	0.167 (0.139)	0.484	2.995 (2.309)	0.651	0.054 (0.415)	0.541	0.617* (0.281)	0.239	3.993 (3.107)	0.833
b09	0.280* (0.111)	-	-13.201*** (2.752)	-	-4.303** (1.596)	-	12.309*** (2.225)	-	0.540 (6.206)	-
b09*SPA	-0.191 (0.141)	0.118	4.035 (3.460)	0.275	-0.763 (1.995)	0.177	-5.197+ (2.712)	0.127	1.755 (7.949)	0.453
b10	0.709*** (0.114)	-	-12.747*** (2.752)	-	-0.608 (0.471)	-	12.873*** (2.250)	-	1.404 (6.017)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
b10*SPA	0.015 (0.146)	0.329	4.243 (3.460)	0.487	-0.148 (0.589)	0.388	-5.049+ (2.744)	0.082	3.236 (7.912)	0.665
b11	0.008 (0.110)	-	-1.343*** (0.295)	-	-2.853** (1.000)	-	-1.210*** (0.250)	-	-1.656+ (0.965)	-
b11*SPA	0.227 (0.140)	0.545	0.646+ (0.372)	0.696	-0.130 (1.250)	0.603	0.735* (0.308)	0.301	-0.095 (1.244)	0.872
b12	0.008 (0.111)	-	-13.52*** (2.752)	-	-9.826** (3.418)	-	-1.208*** (0.250)	-	-8.227+ (4.531)	-
b12*SPA	0.388** (0.143)	0.709	4.658 (3.460)	0.911	-0.837 (4.272)	0.772	0.897** (0.309)	0.467	2.728 (5.589)	1.113
b13	-0.178 (0.109)	-	-4.226*** (0.832)	-	-1.077** (0.331)	-	-1.397*** (0.250)	-	-2.937** (0.936)	-
b13*SPA	0.236+ (0.139)	0.554	1.505 (1.046)	0.710	0.124 (0.415)	0.612	0.744* (0.308)	0.311	1.891 (1.241)	0.889
b14	-0.529*** (0.109)	-	-8.634*** (1.654)	-	-5.787** (1.830)	-	12.159*** (2.345)	-	2.623 (5.272)	-
b14*SPA	0.235+ (0.139)	0.553	2.782 (2.079)	0.720	-0.422 (2.288)	0.611	-5.046+ (2.859)	0.305	-0.582 (6.612)	0.893
b15	-0.012 (0.110)	-	-4.961*** (1.014)	-	-0.222+ (0.132)	-	-1.231*** (0.251)	-	-3.134** (1.191)	-
b15*SPA	0.045 (0.140)	0.359	1.598 (1.275)	0.518	0.018 (0.167)	0.416	0.552+ (0.308)	0.115	2.372 (1.649)	0.696
b16	0.158 (0.110)	-	-9.287*** (1.928)	-	-7.446** (2.645)	-	-1.059*** (0.251)	-	-5.859+ (3.296)	-
b16*SPA	0.038 (0.141)	0.352	3.005 (2.424)	0.517	-0.912 (3.306)	0.411	0.545+ (0.308)	0.107	1.422 (4.056)	0.699
b17	0.863*** (0.116)	-	-12.136*** (2.661)	-	-22.502** (8.124)	-	38.168*** (6.894)	-	14.513 (15.259)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
b17*SPA	-0.321* (0.146)	0.014	3.749 (3.345)	0.127	-3.249 (10.154)	0.039	-15.843+ (8.405)	0.251	-10.641 (18.939)	0.325
b18	-0.006 (0.110)	-	-2.255*** (0.471)	-	-5.924** (2.059)	-	-1.224*** (0.250)	-	-2.408 (2.054)	-
b18*SPA	0.056 (0.140)	0.370	0.762 (0.592)	0.527	-0.684 (2.574)	0.428	0.563+ (0.308)	0.126	-1.079 (2.657)	0.703
b19	-0.053 (0.110)	-	-0.054 (0.108)	-	-0.263* (0.132)	-	-1.271*** (0.250)	-	-0.800** (0.307)	-
b19*SPA	-0.064 (0.139)	0.248	-0.060 (0.137)	0.406	-0.090 (0.166)	0.306	0.443 (0.308)	0.003	0.077 (0.394)	0.581
b20	0.177 (0.110)	-	-2.528*** (0.561)	-	-0.934* (0.401)	-	-1.041*** (0.251)	-	-1.960** (0.625)	-
b20*SPA	-0.189 (0.140)	0.120	0.663 (0.705)	0.282	-0.328 (0.502)	0.178	0.317 (0.308)	0.125	0.790 (0.789)	0.458
b21	0.061 (0.110)	-	-4.441*** (0.924)	-	-24.104** (8.398)	-	-1.157*** (0.250)	-	-5.263 (8.645)	-
b21*SPA	-0.058 (0.140)	0.254	1.357 (1.161)	0.415	-3.077 (10.496)	0.315	0.449 (0.308)	0.010	-7.441 (11.372)	0.593
b22	0.360*** (0.111)	-	-2.349*** (0.561)	-	0.148 (0.133)	-	-0.856*** (0.251)	-	-1.690** (0.599)	-
b22*SPA	-0.014 (0.142)	0.299	0.837 (0.706)	0.460	-0.040 (0.169)	0.357	0.493 (0.309)	0.054	1.321 (0.822)	0.635
b23	-0.530*** (0.112)	-	-10.440*** (2.019)	-	-6.482** (2.071)	-	-1.166*** (0.162)	-	-6.285* (3.086)	-
b23*SPA	0.773*** (0.144)	1.102	3.897 (2.538)	1.284	0.031 (2.589)	1.162	1.041*** (0.203)	0.859	2.934 (3.836)	1.479
b24	-0.193+ (0.109)	-	-9.647*** (1.928)	-	-12.311** (4.212)	-	-0.833*** (0.161)	-	-6.336 (4.670)	-

Effect	Base model Estimate (SE)	Base model Effect Size	LEX Predictor Estimate (SE)	LEX Predictor Effect Size	NP Predictor Estimate (SE)	NP Predictor Effect Size	RC Predictor Estimate (SE)	RC Predictor Effect Size	All Predictors Estimate (SE)	All Predictors Effect Size
b24*SPA	0.070 (0.139)	0.385	3.040 (2.424)	0.553	-1.444 (5.265)	0.444	0.337+ (0.200)	0.141	-0.440 (5.801)	0.735
b25	-0.004 (0.110)	-	-4.055*** (0.832)	-	-2.647** (0.925)	-	-1.223*** (0.250)	-	-2.940* (1.175)	-
b25*SPA	0.251+ (0.140)	0.569	1.520 (1.046)	0.726	-0.080 (1.156)	0.627	0.759* (0.308)	0.326	1.213 (1.448)	0.903
b26	0.166 (0.110)	-	-1.641*** (0.383)	-	-0.045 (0.132)	-	-1.052*** (0.251)	-	-1.451*** (0.385)	-
b26*SPA	0.093 (0.141)	0.408	0.658 (0.482)	0.564	0.066 (0.168)	0.465	0.600+ (0.308)	0.164	1.027* (0.518)	0.740
b27	-0.528*** (0.109)	-	-13.593*** (2.660)	-	-3.211*** (0.939)	-	-1.459*** (0.204)	-	-7.636* (3.408)	-
b27*SPA	0.080 (0.138)	0.395	4.182 (3.344)	0.569	-0.257 (1.174)	0.451	0.467+ (0.251)	0.151	4.964 (4.526)	0.754
b28	0.149 (0.110)	-	-6.601*** (1.379)	-	-15.447** (5.421)	-	50.797*** (9.354)	-	25.545 (15.601)	-
b28*SPA	0.106 (0.141)	0.422	2.226 (1.734)	0.584	-1.842 (6.775)	0.482	-20.976+ (11.404)	0.168	-12.341 (19.374)	0.746
b29	-0.591*** (0.110)	-	-2.382*** (0.383)	-	-3.443*** (0.998)	-	-1.813*** (0.251)	-	-2.460* (0.972)	-
b29*SPA	0.197 (0.139)	0.514	0.759 (0.482)	0.667	-0.161 (1.247)	0.571	0.705* (0.308)	0.271	0.079 (1.230)	0.843
b30	-0.431*** (0.109)	-	-7.635*** (1.471)	-	-3.283*** (0.997)	-	-1.651*** (0.250)	-	-4.907** (1.886)	-
b30*SPA	0.130 (0.139)	0.446	2.393 (1.850)	0.610	-0.228 (1.247)	0.502	0.638* (0.308)	0.202	2.409 (2.404)	0.792
b31	-0.342** (0.109)	-	-6.193*** (1.197)	-	-1.240*** (0.331)	-	-1.562*** (0.250)	-	-3.968** (1.411)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
b31*SPA	0.245+ (0.139)	0.563	2.082 (1.505)	0.724	0.131 (0.414)	0.619	0.753* (0.308)	0.320	2.701 (1.902)	0.905
b32	-0.441*** (0.112)	-	-9.003*** (1.745)	-	-7.659** (2.510)	-	-1.659*** (0.251)	-	-6.027* (3.053)	-
b32*SPA	0.708*** (0.144)	1.036	3.406 (2.193)	1.213	-0.191 (3.137)	1.097	1.219*** (0.309)	0.795	1.876 (3.756)	1.410
b33	0.732*** (0.114)	-	-1.986*** (0.561)	-	0.518*** (0.135)	-	-0.483+ (0.252)	-	-1.328* (0.599)	-
b33*SPA	-0.149 (0.146)	0.161	0.706 (0.706)	0.326	-0.174 (0.172)	0.220	0.357 (0.311)	0.084	1.191 (0.823)	0.503
b34	-0.043 (0.110)	-	-2.742*** (0.561)	-	-0.942** (0.331)	-	-0.973*** (0.204)	-	-1.982** (0.633)	-
b34*SPA	-0.244+ (0.139)	0.064	0.607 (0.705)	0.225	-0.356 (0.414)	0.122	0.142 (0.251)	0.181	0.766 (0.816)	0.402
b35	0.176 (0.111)	-	-9.715*** (2.019)	-	-2.242** (0.847)	-	-1.041*** (0.251)	-	-5.529* (2.524)	-
b35*SPA	0.033 (0.141)	0.347	3.139 (2.538)	0.511	-0.269 (1.059)	0.405	0.539+ (0.308)	0.101	3.671 (3.325)	0.690
b36	0.208+ (0.110)	-	-8.785*** (1.836)	-	-12.618** (4.459)	-	38.901*** (7.147)	-	17.675 (12.686)	-
b36*SPA	0.092 (0.141)	0.407	2.918 (2.309)	0.572	-1.511 (5.573)	0.466	-16.014+ (8.713)	0.156	-8.080 (15.773)	0.739
b37	-0.178 (0.109)	-	-5.578*** (1.106)	-	-3.284** (1.085)	-	-0.819*** (0.161)	-	-3.463* (1.614)	-
b37*SPA	-0.030 (0.139)	0.283	1.666 (1.390)	0.444	-0.418 (1.356)	0.341	0.237 (0.200)	0.039	1.242 (2.008)	0.621
b38	-0.046 (0.110)	-	-5.897*** (1.197)	-	-4.867** (1.679)	-	-0.046 (0.110)	-	-3.339 (2.182)	-
b38*SPA	0.139 (0.140)	0.455	1.976 (1.505)	0.616	-0.464 (2.099)	0.513	0.140 (0.140)	0.211	0.814 (2.690)	0.795

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
b39	-0.532*** (0.109)	-	-2.771*** (0.471)	-	-0.530*** (0.109)	-	-1.753*** (0.250)	-	-2.328*** (0.508)	-
b39*SPA	-0.058 (0.138)	0.254	0.648 (0.592)	0.411	-0.059 (0.138)	0.311	0.450 (0.308)	0.011	1.160 (0.706)	0.589
b40	-0.105 (0.109)	-	-5.505*** (1.105)	-	-15.963** (5.511)	-	-0.167 (0.110)	-	-4.332 (5.746)	-
b40*SPA	0.017 (0.139)	0.331	1.713 (1.390)	0.492	-1.964 (6.888)	0.390	0.043 (0.140)	0.087	-3.879 (7.402)	0.671
b41	-0.012 (0.110)	-	-4.513*** (0.924)	-	-1.768** (0.620)	-	-0.941*** (0.204)	-	-2.902* (1.154)	-
b41*SPA	-0.141 (0.139)	0.169	1.275 (1.161)	0.331	-0.360 (0.775)	0.228	0.246 (0.251)	0.075	1.324 (1.472)	0.508
b42	-0.156 (0.109)	-	-11.857*** (2.386)	-	-24.794** (8.561)	-	-1.375*** (0.250)	-	-8.983 (8.933)	-
b42*SPA	0.263+ (0.140)	0.582	3.944 (2.999)	0.757	-2.811 (10.701)	0.646	0.771* (0.308)	0.338	-4.108 (11.362)	0.947
b43	-0.156 (0.110)	-	-7.358*** (1.471)	-	-5.617** (1.901)	-	12.165*** (2.278)	-	3.189 (5.049)	-
b43*SPA	-0.032 (0.139)	0.281	2.230 (1.850)	0.444	-0.714 (2.376)	0.339	-5.159+ (2.777)	0.035	-1.261 (6.311)	0.619
b44	-0.365** (0.112)	-	-14.798*** (2.935)	-	-8.292** (2.757)	-	-0.071 (0.125)	-	-7.948+ (4.548)	-
b44*SPA	0.481*** (0.143)	0.804	5.035 (3.690)	1.008	-0.507 (3.445)	0.865	0.362* (0.158)	0.561	3.702 (5.688)	1.210
b45	-2.627*** (0.120)	-	-11.181*** (1.745)	-	-9.492*** (2.389)	-	-2.974*** (0.136)	-	-7.641* (3.089)	-

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Estimate (SE)	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**	**Estimate (SE)**	**Effect Size**
b45*SPA	1.465*** (0.149)	1.808	4.137+ (2.194)	1.959	0.607 (2.985)	1.867	1.608*** (0.168)	1.561	2.557 (3.807)	2.124
delta1	4.182*** (0.041)	-	4.209*** (0.042)	-	4.182*** (0.041)	-	4.188*** (0.041)	-	4.235*** (0.042)	-
delta1*SPA	0.038 (0.057)	-	0.022 (0.058)	-	0.038 (0.057)	-	0.035 (0.058)	-	0.008 (0.058)	-
delta2	5.961*** (0.089)	-	6.002*** (0.090)	-	5.962*** (0.089)	-	5.969*** (0.090)	-	6.043*** (0.090)	-
delta2*SPA	0.171 (0.135)	-	0.145 (0.136)	-	0.170 (0.135)	-	0.166 (0.135)	-	0.121 (0.136)	-
delta3	7.138*** (0.157)	-	7.185*** (0.158)	-	7.139*** (0.157)	-	7.148*** (0.157)	-	7.233*** (0.159)	-
delta3*SPA	0.205 (0.243)	-	0.174 (0.243)	-	0.204 (0.243)	-	0.200 (0.243)	-	0.145 (0.244)	-
Intercept Variance	0.462		0.463		0.455		0.448		0.465	
LEX Variance	-		0.022		-		-		0.037	
NP Variance	-		-		0.001		-		0.001	
RC Variance	-		-		-		0.004		0.020	
Intercept*Feature Covariance	-		0.087		0.017		-0.037		See Table G30	

Note: + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate DIF significance and direction: dark blue for substantial DIF favoring the reference group, light blue for moderate DIF favoring the reference group, dark brown for substantial DIF favoring the focal group, and light brown for moderate DIF favoring the focal group.

Table G29.

OTHvSPA Models' Adjusted DIF Estimates – Biology Assessment

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b01*SPA	0.304 (0.139)	[0.032, 0.576]	0.473 (4.151)	[-7.663, 8.609]	0.361 (3.373)	[-6.250, 6.972]	0.058 (8.682)	[-16.959, 17.075]	0.640 (16.677)	[-32.047, 33.327]
b02*SPA	-	-	-	-	-	-	-	-	-	-
b03*SPA	0.144 (0.140)	[-0.130, 0.418]	0.302 (1.964)	[-3.547, 4.151]	0.202 (3.245)	[-6.158, 6.562]	-0.097* (0.308)	[-0.700, 0.507]	0.476 (3.683)	[-6.743, 7.695]
b04*SPA	0.109 (0.140)	[-0.165, 0.383]	0.267 (2.194)	[-4.034, 4.567]	0.166 (1.268)	[-2.319, 2.652]	-0.132* (0.308)	[-0.735, 0.472]	0.440 (2.860)	[-5.166, 6.046]
b05*SPA	0.179 (0.140)	[-0.095, 0.453]	0.336 (1.964)	[-3.513, 4.185]	0.236 (0.791)	[-1.315, 1.786]	-0.062* (0.308)	[-0.665, 0.542]	0.510 (2.519)	[-4.427, 5.447]
b06*SPA	0.306 (0.138)	[0.036, 0.576]	0.460* (0.819)	[-1.145, 2.065]	0.363 (3.422)	[-6.344, 7.070]	0.066* (0.307)	[-0.535, 0.668]	0.634 (3.557)	[-6.338, 7.605]
b07*SPA	0.192 (0.138)	[-0.078, 0.462]	0.347* (0.932)	[-1.479, 2.174]	0.249 (0.966)	[-1.645, 2.142]	-0.048* (0.307)	[-0.649, 0.554]	0.521 (1.230)	[-1.889, 2.932]
b08*SPA	0.474 (0.139)	[0.202, 0.746]	0.638 (2.309)	[-3.888, 5.163]	0.530 (0.415)	[-0.283, 1.343]	0.234* (0.281)	[-0.316, 0.785]	0.816 (3.107)	[-5.274, 6.906]
b09*SPA	0.116 (0.141)	[-0.160, 0.392]	0.269 (3.460)	[-6.512, 7.051]	0.174 (1.995)	[-3.737, 4.084]	-0.124 (2.712)	[-5.44, 5.191]	0.444 (7.949)	[-15.136, 16.024]
b10*SPA	0.322 (0.146)	[0.036, 0.608]	0.477 (3.460)	[-6.304, 7.259]	0.380 (0.589)	[-0.775, 1.534]	0.081 (2.744)	[-5.298, 5.459]	0.652 (7.912)	[-14.856, 16.159]
b11*SPA	0.534 (0.140)	[0.260, 0.808]	0.682* (0.372)	[-0.047, 1.411]	0.591 (1.25)	[-1.859, 3.041]	0.295* (0.308)	[-0.308, 0.899]	0.854 (1.244)	[-1.584, 3.292]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**	**Adj. Estimate (SE)**	**95% CI**
b12*SPA	0.695* (0.143)	[0.415, 0.975]	0.892 (3.460)	[-5.889, 7.674]	0.756 (4.272)	[-7.617, 9.129]	0.457* (0.309)	[-0.148, 1.063]	1.090 (5.589)	[-9.864, 12.045]
b13*SPA	0.543 (0.139)	[0.271, 0.815]	0.696* (1.046)	[-1.354, 2.746]	0.600 (0.415)	[-0.213, 1.413]	0.304* (0.308)	[-0.299, 0.908]	0.871 (1.241)	[-1.562, 3.303]
b14*SPA	0.542 (0.139)	[0.270, 0.814]	0.706 (2.079)	[-3.369, 4.781]	0.599 (2.288)	[-3.886, 5.083]	0.299 (2.859)	[-5.305, 5.903]	0.875 (6.612)	[-12.084, 13.835]
b15*SPA	0.352 (0.140)	[0.078, 0.626]	0.508 (1.275)	[-1.991, 3.007]	0.408* (0.167)	[0.080, 0.735]	0.112* (0.308)	[-0.491, 0.716]	0.682 (1.649)	[-2.550, 3.914]
b16*SPA	0.345 (0.141)	[0.069, 0.621]	0.506 (2.424)	[-4.245, 5.257]	0.402 (3.306)	[-6.077, 6.882]	0.105* (0.308)	[-0.498, 0.709]	0.685 (4.056)	[-7.265, 8.634]
b17*SPA	-0.014* (0.146)	[-0.300, 0.272]	0.124 (3.345)	[-6.432, 6.681]	0.038 (10.154)	[-19.864, 19.940]	-0.246 (8.405)	[-16.720, 16.228]	0.319 (18.939)	[-36.802, 37.439]
b18*SPA	0.363 (0.140)	[0.089, 0.637]	0.517* (0.592)	[-0.644, 1.677]	0.420 (2.574)	[-4.626, 5.465]	0.123* (0.308)	[-0.480, 0.727]	0.689 (2.657)	[-4.519, 5.896]
b19*SPA	0.243 (0.139)	[-0.029, 0.515]	0.398* (0.137)	[0.130, 0.667]	0.300* (0.166)	[-0.026, 0.625]	0.003* (0.308)	[-0.600, 0.607]	0.569 (0.394)	[-0.203, 1.341]
b20*SPA	0.118 (0.140)	[-0.156, 0.392]	0.276* (0.705)	[-1.105, 1.658]	0.174 (0.502)	[-0.810, 1.158]	-0.123* (0.308)	[-0.726, 0.481]	0.449 (0.789)	[-1.098, 1.995]
b21*SPA	0.249 (0.140)	[-0.025, 0.523]	0.407 (1.161)	[-1.869, 2.682]	0.308 (10.496)	[-20.264, 20.881]	0.009* (0.308)	[-0.594, 0.613]	0.581 (11.372)	[-21.708, 22.870]
b22*SPA	0.293 (0.142)	[0.015, 0.571]	0.450* (0.706)	[-0.933, 1.834]	0.350* (0.169)	[0.019, 0.681]	0.053* (0.309)	[-0.552, 0.659]	0.622 (0.822)	[-0.989, 2.234]
b23*SPA	1.08* (0.144)	[0.798, 1.362]	1.258 (2.538)	[-3.716, 6.233]	1.139 (2.589)	[-3.936, 6.213]	0.842* (0.203)	[0.444, 1.240]	1.449 (3.836)	[-6.070, 8.967]

Effect	Base model		LEX Predictor		NP Predictor		RC Predictor		All Predictors	
	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI	Adj. Estimate (SE)	95% CI
b24*SPA	0.377 (0.139)	[0.105, 0.649]	0.541 (2.424)	[-4.210, 5.292]	0.435 (5.265)	[-9.885, 10.754]	0.138* (0.200)	[-0.254, 0.530]	0.720 (5.801)	[-10.650, 12.090]
b25*SPA	0.558 (0.140)	[0.284, 0.832]	0.711* (1.046)	[-1.339, 2.761]	0.614 (1.156)	[-1.652, 2.880]	0.319* (0.308)	[-0.284, 0.923]	0.885 (1.448)	[-1.953, 3.723]
b26*SPA	0.400 (0.141)	[0.124, 0.676]	0.553* (0.482)	[-0.392, 1.497]	0.456* (0.168)	[0.127, 0.785]	0.160* (0.308)	[-0.443, 0.764]	0.725 (0.518)	[-0.291, 1.740]
b27*SPA	0.387 (0.138)	[0.117, 0.657]	0.557 (3.344)	[-5.997, 7.112]	0.442 (1.174)	[-1.859, 2.743]	0.148* (0.251)	[-0.344, 0.640]	0.738 (4.526)	[-8.133, 9.609]
b28*SPA	0.413 (0.141)	[0.137, 0.689]	0.572 (1.734)	[-2.826, 3.971]	0.472 (6.775)	[-12.807, 13.751]	0.165 (11.404)	[-22.187, 22.517]	0.731 (19.374)	[-37.242, 38.704]
b29*SPA	0.504 (0.139)	[0.232, 0.776]	0.654* (0.482)	[-0.291, 1.598]	0.559 (1.247)	[-1.885, 3.003]	0.265* (0.308)	[-0.338, 0.869]	0.826 (1.230)	[-1.584, 3.237]
b30*SPA	0.437 (0.139)	[0.165, 0.709]	0.598 (1.850)	[-3.028, 4.224]	0.492 (1.247)	[-1.952, 2.936]	0.198* (0.308)	[-0.405, 0.802]	0.776 (2.404)	[-3.936, 5.487]
b31*SPA	0.552 (0.139)	[0.280, 0.824]	0.709 (1.505)	[-2.240, 3.659]	0.607 (0.414)	[-0.205, 1.418]	0.313* (0.308)	[-0.290, 0.917]	0.887 (1.902)	[-2.841, 4.615]
b32*SPA	1.015* (0.144)	[0.733, 1.297]	1.189 (2.193)	[-3.110, 5.487]	1.075 (3.137)	[-5.073, 7.224]	0.779* (0.309)	[0.174, 1.385]	1.381 (3.756)	[-5.981, 8.743]
b33*SPA	0.158 (0.146)	[-0.128, 0.444]	0.319* (0.706)	[-1.064, 1.703]	0.216* (0.172)	[-0.121, 0.553]	-0.083* (0.311)	[-0.692, 0.527]	0.492 (0.823)	[-1.121, 2.106]
b34*SPA	0.063 (0.139)	[-0.209, 0.335]	0.220* (0.705)	[-1.161, 1.602]	0.120* (0.414)	[-0.692, 0.931]	-0.177* (0.251)	[-0.669, 0.315]	0.394 (0.816)	[-1.206, 1.993]
b35*SPA	0.340 (0.141)	[0.064, 0.616]	0.500 (2.538)	[-4.474, 5.475]	0.397 (1.059)	[-1.679, 2.472]	0.099* (0.308)	[-0.504, 0.703]	0.676 (3.325)	[-5.841, 7.193]

Effect	Base model Adj. Estimate (SE)	Base model 95% CI	LEX Predictor Adj. Estimate (SE)	LEX Predictor 95% CI	NP Predictor Adj. Estimate (SE)	NP Predictor 95% CI	RC Predictor Adj. Estimate (SE)	RC Predictor 95% CI	All Predictors Adj. Estimate (SE)	All Predictors 95% CI
b36*SPA	0.399 (0.141)	[0.123, 0.675]	0.561 (2.309)	[-3.965, 5.086]	0.457 (5.573)	[-10.466, 11.380]	0.153 (8.713)	[-16.925, 17.230]	0.724 (15.773)	[-30.191, 31.639]
b37*SPA	0.277 (0.139)	[0.005, 0.549]	0.435 (1.390)	[-2.290, 3.159]	0.334 (1.356)	[-2.324, 2.992]	0.038* (0.200)	[-0.354, 0.430]	0.609 (2.008)	[-3.327, 4.544]
b38*SPA	0.446 (0.140)	[0.172, 0.720]	0.603 (1.505)	[-2.346, 3.553]	0.502 (2.099)	[-3.612, 4.616]	0.207* (0.140)	[-0.067, 0.482]	0.779 (2.690)	[-4.493, 6.052]
b39*SPA	0.249 (0.138)	[-0.021, 0.519]	0.403* (0.592)	[-0.758, 1.563]	0.304* (0.138)	[0.034, 0.575]	0.010* (0.308)	[-0.593, 0.614]	0.577 (0.706)	[-0.807, 1.960]
b40*SPA	0.324 (0.139)	[0.052, 0.596]	0.482 (1.390)	[-2.243, 3.206]	0.382 (6.888)	[-13.118, 13.883]	0.085* (0.140)	[-0.190, 0.359]	0.658 (7.402)	[-13.850, 15.166]
b41*SPA	0.166 (0.139)	[-0.106, 0.438]	0.325 (1.161)	[-1.951, 2.600]	0.223 (0.775)	[-1.296, 1.742]	-0.073* (0.251)	[-0.565, 0.419]	0.498 (1.472)	[-2.387, 3.383]
b42*SPA	0.570 (0.14)	[0.296, 0.844]	0.742 (2.999)	[-5.136, 6.620]	0.633 (10.701)	[-20.341, 21.607]	0.331* (0.308)	[-0.272, 0.935]	0.928 (11.362)	[-21.341, 23.198]
b43*SPA	0.275 (0.139)	[0.003, 0.547]	0.435 (1.850)	[-3.191, 4.061]	0.332 (2.376)	[-4.325, 4.989]	0.034 (2.777)	[-5.409, 5.477]	0.607 (6.311)	[-11.763, 12.976]
b44*SPA	0.788* (0.143)	[0.508, 1.068]	0.988 (3.690)	[-6.244, 8.220]	0.848 (3.445)	[-5.904, 7.600]	0.550* (0.158)	[0.240, 0.859]	1.186 (5.688)	[-9.963, 12.334]
b45*SPA	1.772* (0.149)	[1.480, 2.064]	1.920 (2.194)	[-2.381, 6.220]	1.829 (2.985)	[-4.021, 7.680]	1.529* (0.168)	[1.200, 1.859]	2.081 (3.807)	[-5.381, 9.542]

Note: * denotes the adjusted DIF estimate is outside of the confidence interval (CI).

Table G30.

Covariance Matrices for All Predictors Models – Biology Assessment

Comparison Group	Component	Intercept	LEX	NP	RC
EPvEB	Intercept	1.019	**0.924**	**0.756**	**-0.990**
	LEX	0.211	0.051	**0.861**	**-0.943**
	NP	0.053	0.014	0.005	**-0.803**
	RC	-0.198	-0.042	-0.011	0.039
EPvSTEB	Intercept	1.042	**0.924**	**0.746**	**-0.991**
	LEX	0.215	0.052	**0.869**	**-0.944**
	NP	0.054	0.014	0.005	**-0.795**
	RC	-0.201	-0.043	-0.011	0.040
EPvLTEB	Intercept	1.063	**0.922**	**0.761**	**-0.989**
	LEX	0.216	0.052	**0.874**	**-0.945**
	NP	0.059	0.015	0.006	**-0.813**
	RC	-0.205	-0.043	-0.012	0.041
STEBvLTEB	Intercept	0.501	**0.903**	**0.266**	**-0.993**
	LEX	0.125	0.038	**-0.072**	**-0.902**
	NP	0.007	-0.001	0.002	**-0.275**
	RC	-0.102	-0.026	-0.002	0.021
EPvSPA	Intercept	1.033	**0.922**	**0.752**	**-0.990**
	LEX	0.212	0.051	**0.863**	**-0.941**
	NP	0.055	0.014	0.005	**-0.799**
	RC	-0.200	-0.042	-0.011	0.040
EPvOTH	Intercept	1.068	**0.924**	**0.751**	**-0.989**
	LEX	0.219	0.053	**0.868**	**-0.946**
	NP	0.058	0.015	0.006	**-0.805**
	RC	-0.206	-0.044	-0.012	0.040
OTHvSPA	Intercept	0.465	**0.887**	**0.225**	**-0.994**
	LEX	0.117	0.037	**-0.134**	**-0.885**
	NP	0.005	-0.001	0.001	**-0.234**
	RC	-0.095	-0.024	-0.001	0.020

Note: Variances are on the diagonal, covariances are in the lower triangle, and correlations are in the upper triangle in bold.